Aalto University
School of Electrical Engineering
Degree Programme in Electronics and Electrical Engineering

Thomas Svedström

Aural Multitasking in Personal Media Devices

Master's Thesis
Espoo, July 23, 2014

Supervisor: Professor Vesa Välimäki, Aalto University
Instructor: Aki Härmä D.Sc. (Tech.)

Aalto University
School of Electrical Engineering
Degree Programme in Electronics and Electrical Engineering

ABSTRACT OF
MASTER'S THESIS

Author:	Thomas Svedström		
Title:	Aural Multitasking in Personal Media Devices		
Date:	July 23, 2014	**Pages:**	112
Professorship:	S-89	**Code:**	S-89
Supervisor:	Professor Vesa Välimäki		
Instructor:	Aki Härmä D.Sc. (Tech.)		

The use of personal media devices (PMDs) in traffic can lead to safety critical situations. This is due to divided visual attention between the device and the interface. This thesis considers the use of auditory interfaces for multitasking in PMDs. Aural multitasking refers to performing several simultaneous tasks by using sound as the primary display modality. In order to create such an eyes-free multitasking interface, the problems of presenting information from various sound sources and issues regarding the interaction must be solved.

This thesis consists of three distinct topics. The first topic presents a gesture controller for auditory interfaces. The controller uses acoustic classification to recognize four tactile gestures and it can be operated for example through a pocket. The second topic presents a multilayer auditory interface. The multilayer interface incorporates ideas from ambient displays and creates a personal, layered, soundscape that enables auditory attention managing. The method divides the information and tasks into foreground and background streams according to their priorities. The last topic presents a rapid head-related transfer function (HRTF) personalization method for PMD usage. The method is implemented as an auditory game and it does not require additional accessories besides the headphones.

Keywords:	Auditory interfaces, eyes-free interaction, acoustics, acoustic signal processing, spatial sound, sonification
Language:	English

Aalto-yliopisto
Sahkotekniikan korkeakoulu
Elektroniikan ja sähkötekniikan tutkinto-ohjelma

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä:	Thomas Svedström		
Työn nimi:	Äänenvarainen monikäyttö henkilökohtaisissa medialaitteissa		
Päiväys:	23. heinäkuuta 2014	**Sivumäärä:**	112
Professuuri:	S-89	**Koodi:**	S-89
Valvoja:	Professori Vesa Välimäki		
Ohjaaja:	Tekniikan tohtori Aki Härmä		

Henkilökohtaisten medialaitteiden (personal media device, PMD) käyttö liikenteessä saattaa johtaa onnettomuuksiin. Tämä johtuu kyseisten laitteiden käytön aikana tapahtuvasta visuaalisen huomiokyvyn jakamisesta laitteen ja ympäristön välillä. Tämä diplomityö käsittelee äänenvaraista monikäyttöä (auditory multitasking) PMD-laitteissa käyttäen lähtökohtanaan äänikäyttöliittymiä. Äänenvarainen monikäyttö viittaa useiden samanaikaisten tehtävien suorittamiseen käyttäen ääntä ensisijaisena modaliteettina. Jotta tähän tavoitteiseen päästäisiin, on ratkaistava useita perustavanlaatuisia ongelmia monilähteisen ääni-informaation esittämiseen ja interaktioon liittyen.

Tämä diplomityö koostuu kolmesta aiheesta. Ensimmäinen aihe esittelee eleisiin perustuvan ohjaustavan äänikäyttöliittymille. Ohjain käyttää äänentunnistusta neljän haptisen eleen luokitteluun. Tästä johtuen ohjainta voidaan käyttää esimerkiksi taskun läpi. Toinen aihepiiri esittelee monikerroksisen äänikäyttöliittymän, joka hyödyntää ns. ympäristöön sulautuvien näyttöjen (ambient display) ideoita ja luo henkilökohtaisen, kerrostetun äänimaiseman. Tarkoituksena on luoda äänimaisema, jossa käyttäjä pystyy keskittämään huomiokykynsä haluamaansa äänivirtaan. Kyseisessä toteutuksessa äänilähteet jaotellaan etu- ja taustakerroksiin niiden prioriteettien perusteella. Viimeinen aihe esittelee nopean head-related transfer function -pohjaisen (HRTF) tilaäänijärjestelmän personalisointimetodin. Metodi voidaan toteuttaa äänipelinä ja se ei vaadi kuulokkeiden lisäksi erillisiä laitteita.

Asiasanat:	Äänikäyttöliittymät, katseeton vuorovaikutus, akustiikka, akustinen signaalinkäsittely, tilaääni, sonifikaatio
Kieli:	Englanti

Acknowledgements

The research for the thesis was performed in Philips Research Eindhoven in the Netherlands.

I would like to express my great gratitude to my instructor Aki Härmä D.Sc. (Tech.) for offering me this opportunity. I had the freedom to present my ideas to Härmä, who then helped me to develop them even further in a very intellectual fashion. I was pleased to work under his supervision.

I would also like to thank my supervisor professor Vesa Välimäki for his help and feedback on the thesis and providing valuable information on the final phase of the writing part. Professor Välimäki has also inspired me during my master's studies by giving some very interesting courses and lectures. These courses kept me highly motivated during my time as a student at the department of Signal Processing and Acoustics. That said, I would also like to thank all the lecturers, course assistants and fellow students.

I would also like to thank my family, friends and especially Mailis for providing support, kindness and love during this period.

Helsinki, July 23, 2014

Thomas Svedström

Abbreviations and Acronyms

BRIR	Binaural impulse response
CLT	Cognitive load theory
CRM	Coordinate Response Measurement
D/R	Direct-to-reverberant ratio
ECG	Electrocardiogram
GUI	Graphical user interface
HAT	Head and torso
HED/UT	Hedonic Utility scale
HRIR	Head-related impulse response
HRTF	Head-related transfer function
IID	Interaural intensity difference
ILD	Interaural level difference
IPD	Interaural phase difference
ITD	Interaural time difference
JND	Just noticeable difference
MAA	Minimum audible angle
MBS	Model based sonification
NASA-TLX	NASA Task load index
OSM	Off-screen model
PMD	Personal media device
PMS	Parameter mapping sonification
PRTF	Pinna-related transfer function
RMS	Root mean square
SI	Speech intelligibility
TTS	Text-to-speech
UCD	User centered design
UI	User interface
VAD	Virtual auditory display
WISP	Weakly Intrusive Ambient Soundscape for Intuitive State Perception

Contents

Chapter 1

Introduction

Personal mobile devices (PMD) have enabled the possibility of being constantly connected to social networks, having an instant access to various online services, reading news and emails and having a whole media collection in our pocket. PMDs have proven to be helpful in many cases as the user can peform various tasks regardless of their own physical location. The growth in the use of smartphones has been exponential worldwide and it is predicted to become an everyday object worldwide [1].

The usage of a cellphones in traffic leads to safety critical situations [2, 3]. This is due to the limitations of human attention capabilities, which include both sensory and cognitive factors. For example, the distraction caused by interfaces that require visual attention (i.e. PMDs and music players) in a car is the major contributor in automobile crashes [4, 5, 6]. Furthermore, the pedestrians have an increased risk of getting hit by a car while using PMDs as they tend to look at the device instead of paying attention to the traffic [2]. The visual and cognitive distractions caused by the PMD reduce the situation awareness and increases reaction times and unsafe behavior [3]. The divided visual attention between the PMD and the surrounding environment is fragmented into short bursts that have the duration of $4 - 8$ seconds [7, 8].

1.1 Aim of the thesis

As majority of the distractions caused by the PMDs are visual, this thesis aims to find design concepts that enable PMD usage based solely on the auditory modality. As the users will have their eyes free, the visual inattentional blindness is reduced which results in increased traffic safety.

The interface mapping from the visual domain to auditory is not a straightforward process. How to, for example, attain same information level as a

visual interface can provide? On the other hand, are the two modalities so different by nature that totally different objectives should be considered? Also, what are the user's expectations and what are the tasks they are performing? Furthermore, as the screen is not used, the PMD does not need to be held in hand - it can be virtually located anywhere, for example inside a pocket.

Rather than going into an exact interface design, this thesis presents design ideas and concepts that support aural multitasking in a PMD.

1.2 Workflow

The workflow followed a user centered design (UCD) process. In UCD [9] the user is involved in the planning and prototyping stages of a product development. In the current study, the outcome of the planning stage was three distinct research topics. The workflow of the planning stage consisted of five steps and is presented in Figure 1.1.

First, a group of hypothetical consumers were interviewed about their PMD usage habits. Most subjects reported that they use PMDs for sending and receiving messages, reading news, listening to music, playing games and using lifestyle such as the sports and fitness applications.

Second, five hypothetical consumer insights were formulated based on the interviews. The insights were written in a form of a general consumer wish. Each insight included a usability problem that varied from eyes-free interaction to a type of ambient awareness. The insights are presented in Appendix A.

Third, the insights were presented to the representatives of the marketing division. The marketing division chose three topics on which to be consentrated.

Fourth, a brainstorm session was organized to find practical solutions to the chosen topics. The attendees were briefly introduced to basic auditory interface and spatial sound concepts. The group generated a large number of data and propositions for each topic.

Finally, the data gathered from the brainstorm session was processed and organized into three research topics. First topic considers rapid and eyes-free interaction. The second topic considers information presentation from multiple simultaneous sources. Third topic is about head-phone based spatial sound personalization method that is suitable for PMD usage.

After the planning, the three research topics were independently and consecutively implemented. Each implementation was evaluated by conducting a subjective listening test or experiment.

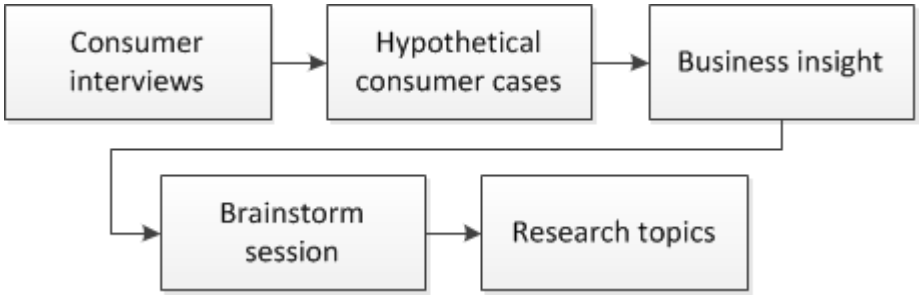

Figure 1.1: The workflow of the planning stage.

1.3 Organization of the thesis

The thesis is divided into seven chapters. Chapter 2 introduces the theory related to spatial sound. Chapter 3 presents key concepts of the auditory interfaces. Chapter 4 presents a gesture based controller for eyes-free interaction. Chapter 5 suggests a multilayer auditory interface for auditory attention managing and multitasking. Chapter 6 presents a rapid HRTF personalisation method. Finally, chapter 7 concludes the thesis.

Chapter 2

Spatial sound

The human auditory system is able to conceive the spatial properties of sounds. These properties include the direction, elevation and distance of a sound source [10, 11]. A listener may also be able to estimate the dimensions of the room and the sound source from the perceived spatial impression [12, 13]. From an evolutionary viewpoint, these abilities have been beneficial in a hostile environment [14].

The properties of spatial hearing have been under intensive research over the last decades. The research has enabled spatial sound reproduction systems that are mostly used in entertainment, i.e. surround sound for movies, music and games, but they have other applications as well. These application areas include teleconferencing, technological aids for visually impaired, clinical use and virtual and augmented reality [15, 16, 17, 18, 19].

2.1 Sound source localization

The source localization is based on binaural listening. For readings concerning the fundamentals of sound source localization, see [10, 20, 21, 22, 11]. A normal human has two ears, a head into which the ears are attached and a torso for the head. Each of these body parts has their own contribution to the sound before entering the ear canals. For example the head attenuates the sound at contralateral ear and the torso causes a so called shoulder bounce for elevated sound sources. The sound varies between the ears in onset time, pressure level and spectral properties. In the literature, time variation is referred to as interaural time difference (ITD) and level difference as interaural level difference (ILD). Based on these variations, the auditory system is able to determine the horizontal direction and elevation of the sound source.

The localization is strongly affected by the anthropometrics [23, 24, 25,

26, 27]. There are countless of smaller and larger differences in the shapes and sizes of pinnae, heads and torsos amongst individuals. Thus, the attenuation and scattering of a sound wave from these body parts varies from person to person, which causes spectral variation. Each person has an auditory system that is adjusted to the cues that are provided by the individual spectral features. However, the system is also able to adapt to changes in the anthropometrics [28]. This can be the case for example after a trauma.

2.1.1 The interaural-polar coordinate system

A polar coordinate system is often used in the context of source localization and spatial sound. The system specifies the direction of an sound source unambiguously by using two angles. The azimuth (θ) corresponds to lateral direction and (ϕ) to elevation. The origin of the system is the midpoint of an interaural axis that is set by a line between the two ears. The vertical plane, often referred to as the median plane, bisects the head into left and right hemispheres. Elevation (ϕ) specifies the rotation around the interaural axis. Azimuth is the angle between a ray to the sound source and the median plane. The coordinate system is presented in Figure 2.1.

2.1.2 Interaural time difference (ITD)

A sound wave rarely ever enters the two ear canals exactly at the same time. Let us consider two cases that are presented in Figure 2.2 and Figure 2.3. It can be observed in the first figure, where a plane wave is arriving directly at the front (ϕ=0°and θ=0°), that the wave propagates equal distance to both ears. Hence, there are no temporal differences between the ears. In the latter figure, the source is located 30° to the right. Here the soundwave has a longer distance to propagate to the left ear than to the right ear. As the sound velocity in air is approximately 340m/s, the wavefront reaches the left ear slightly later. This difference is referred to as the interaural time difference (ITD). The auditory system detects ITDs ranging from 10μs [30] to approximately 700μs.

Directional cues provided by the interaural time difference are frequency dependent. The ITD provides the primary localization cues at low frequencies up to 1000Hz [31]. Above 1kHz there is a transition band up to 2kHz, where the localization is not properly functioning. This is due to the physical dimensions of the head. The average distance between the two ears is approximately 23cm, which corresponds to the wavelength of a 1500Hz sinusoidal signal. Above 1500Hz the wavelength becomes shorter than the distance between the ears. As a result, ITD does not provide localization cues, as it becomes

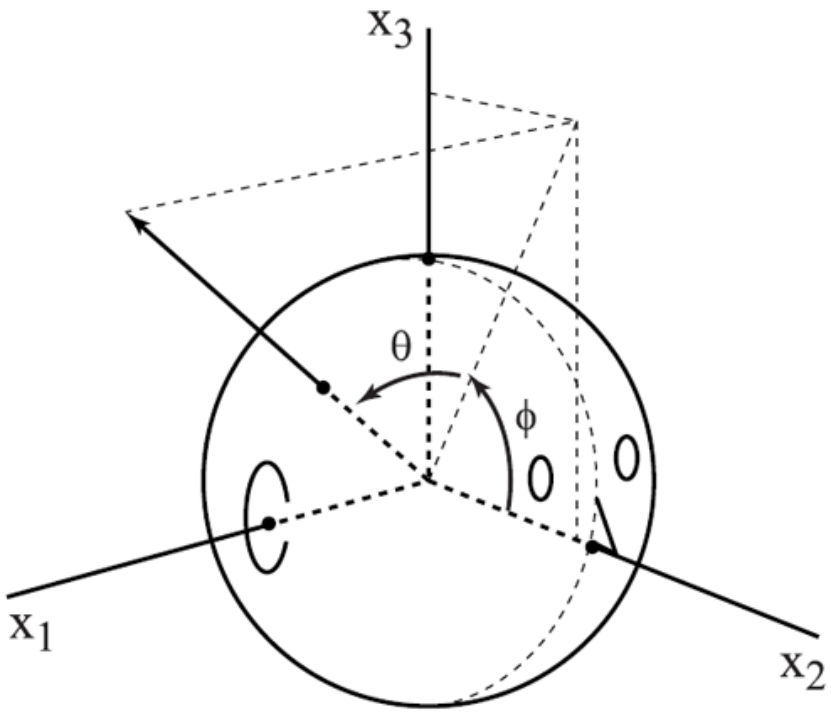

Figure 2.1: The interaural-polar coordinate system [29].

impossible to determine which signal is behind. Also, at very low frequencies the ITD is unable to provide localization cues, because the wavelength is too long compared to the interaural distance. However, the auditory system is able to determine the ITDs of complex signals from the interaural envelope [32].

2.1.3 Interaural level difference (ILD)

There are almost always sound pressure level differences between the ears. Let us again consider the Figure 2.2 and Figure 2.3. It can be inferred, that the sound pressure level is very close to equal at both ears when the source it at front. In Figure 2.3, the right ear is closer to the sound source. The head, as an physical object with a certain mass, volume and internal structure, attenuates the sound wave. This interaural level difference (ILD) can be also referred to as interaural intensity difference (IID).

ILD is more efficient in providing localization cues at high frequencies [22]. ILD becomes more pronounced approximately above 1500Hz. This is due to the fact that the wavelength of the incoming sound is of the same size as

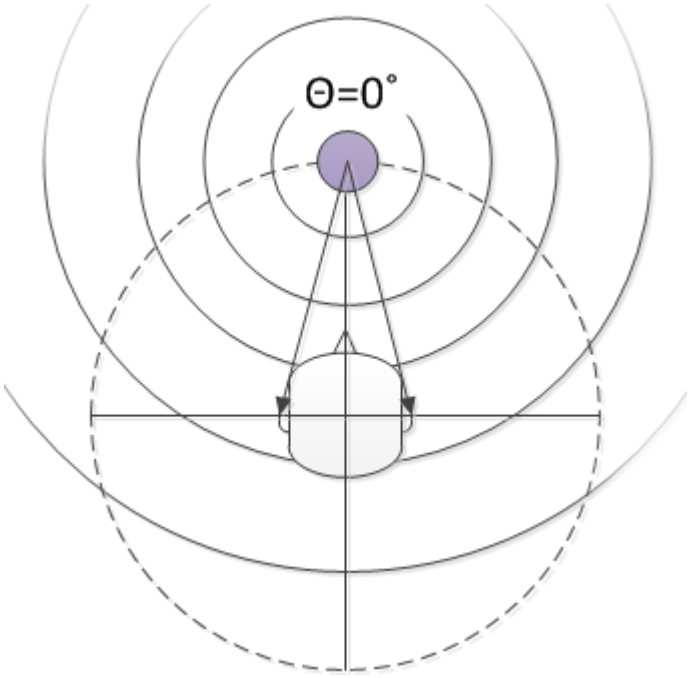

Figure 2.2: The wave propagation of a sound source located at $\theta = 0°$.

the head diameter. Correspondingly, the head is acoustically transparent at low frequencies, and thus there is no detected attenuation between the ears and the localization is based on ITD. Furthermore, ILD provides the most prominent localization cues above 4kHz. For example, at 1kHz the ILD can be 8dB and at 10kHz it may be high as 30dB. The just-noticeable difference (JND) for ILD is approximaltely 0.5dB [30]. JND is a threshold that describes the smallest level that is correctly detected 50% of the time [33].

2.1.4 Issues on localization

Localization is not completely accurate for every sound source. According to the classic duplex theory, ILD and ITD work at complementary frequency ranges - ITD for low and ILD for high frequencies [34, 31]. However, there is a transition band from approximately 1kHz to 4kHz, where neither of the mechanisms are ideal [20]. The highest error rate occurs at 3kHz.

The so called *cone of confusion* [34, 35, 19] is predicted by the duplex theory. The theory assumes the head as a complete sphere. If we draw a cone that has base at $\theta = 90°$, $\phi = 0°$, we notice, that on the circular intersection of the cone, the ILD and ITD are constant. They are thus unable to provide

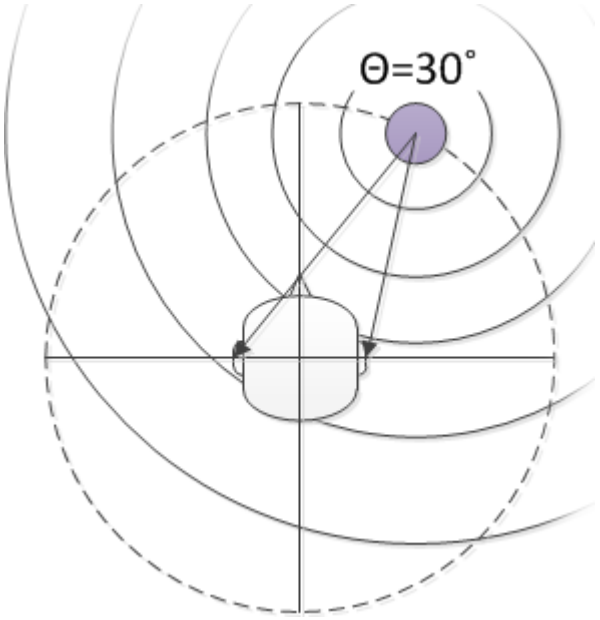

Figure 2.3: The wave propagation of a sound source located at $\theta = 30°$.

unambiguous localization cues. The source may be located at $\theta = -45°$ or at $\theta = 45°$, but the auditory system, according to the duplex theory, would not able to determine whether the source is located at front or back.

The cone of confusion itself is theoretical to some extend. The head is not a regular shaped sphere. Therefore, there are always some slight variations on the ITD and ILD [19]. More importantly, the spectral features that occur due to the sound scattering and diffracting on pinna, head and torso have an significant influence on the localization accuracy [19]. In fact, the spectral balance between frontal and backward directions is the primary cue for front-back discrmimination [36, 37]. Furthermore, the capability to move the head decreases the front-back confusions and increases the localization accuracy [38, 39, 40].

The localization resolution is not evenly distributed. The localization is most accurate at frontal locations and least accurate at the sides. Localization blur expresses the minimum audible angle (MAA) for θ, ϕ and distance. Depending on what kind of excitation signal is used, the localization blur for θ varies from the maximal accuracy of approximately $\pm 1°$ [41] at front to $\pm 10°$ at the sides [11]. It is especially high for 1500Hz sinusoids at the sides. This is again due to the fact that the wavelength is the same size as the head. The localization blur for a 100ms white noise burst is shown in Figure 2.4.

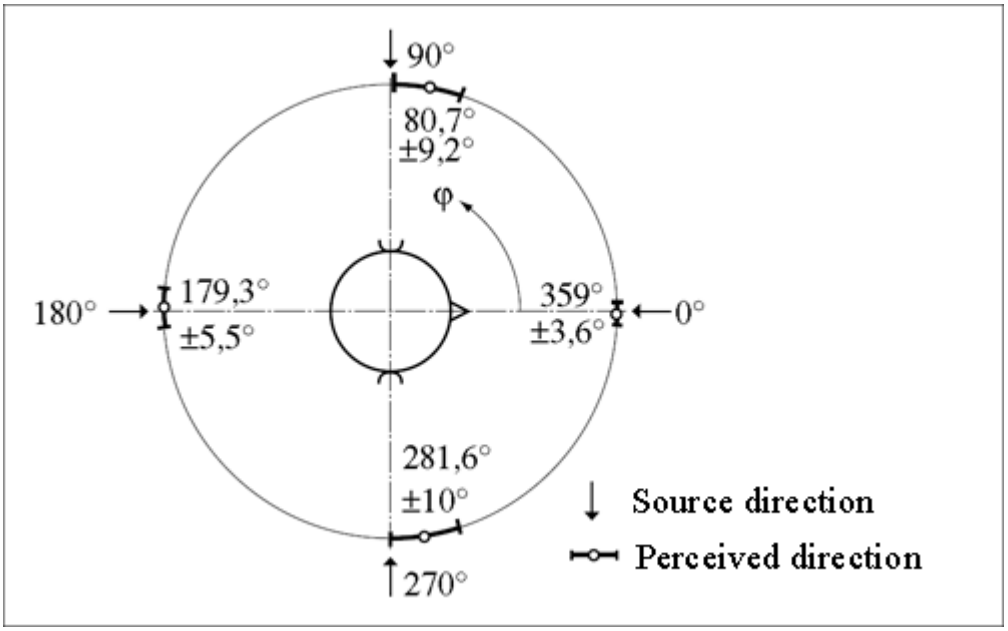

Figure 2.4: Localization blur for 100ms white noise burst. The source positions and corresponding perceived directions are evaluated at four positions. Figure adapted from [11].

2.2 Distance perception

The perception of a sound depends on where the sound source and listener are located and the acoustic environment. The sound propagates in the air and reflects from the surfaces before entering the ear canals of the listener. For example, if the subject is close to the sound source, the sound intensity is higher. Furthermore, if the subject is moving further away, the ratio between direct and reflected sound decreases.

Four acoustic cues are often proposed to explain the capability to perceive the distance of a sound source [42]. The cues include intensity, direct-to-reverberant energy ratio, spectrum and binaural differences. Auditory system combines multiple cues to construct the distance estimate [43]. As the acoustic cues change from an environment to another, learning has an important part in the distance estimation [44, 11]. The accuracy of distance perception is poor compared to direction accuracy [45]. There is a tendency to underestimate the distance of distant sources and to overestimate the distance of sources closer than one meter [43].

2.2.1 Intensity cues

Sound intensity loss follows the inverse-square law as a function of distance. Under ideal free-field conditions using a point source, the intensity decreases 6dB as the distance doubles. Under non-indeal reverberant conditions this loss is always less than 6dB due to the reflections.

Distance perception accuracy is increased if the sound source is familiar [44, 46, 11]. Under non-reverberant conditions, the auditory system is forced to make an assumption about the source power according to the sound intensity at the ear [43]. As is it known how loud a certain sound is on a known distance, the distance can be estimated when the source is at an unknown distance. Distance perception based solely on intensity variations is limited to approximately 10 meters [11]. The term acoustic horizon describes the point, after which changes in the distance in the terms of intensity are no longer perceived [47].

2.2.2 Reverberation

Reverberation is due to the sound reflections from surfaces in a certain space. Room size, surface materials, and the amount and placement of absorbants are the primary factors that determine the reverberation of an environment. Reverberation is often approximated to be a diffuse sound field and therefore the reverberant energy can be assumed to be equally distributed.

As the source distance increases, the ratio between direct and reverberant (D/R) sound parts changes [43]. The direct sound part attenuates according to the inverse-square law, while the reverberant part remains the same according to the diffuse field approximation. As a result, the D/R ratio decreases as the distance increases.

Reverberation seems to provide absolute distance cue that is independend of source power [48, 49]. Furthermore, distance perception is more accurate in reverberant environment than in anechoic [42]. Early reflections and reverberation time provide strong distance cues, as the human listener is able to make assumptions of the size of the space [11].

2.2.3 Spectral properties

The spectral properties have been reported to have an impact on the distance perception. In [50] a click that was low-pass filtered with a cut-off frequency 7.68kHz was consistently judged to be more distant than click that was low-pass filtered at 10.56kHz. However, Little et al. [51] discussed whether these values are too extreme to appear under natural conditions. Furthermore,

their study show that small spectral chances can serve as a relative but not as an absolute cue.

Spectral changes as a function of distance are mainly due to the air absorption and reflections [43]. On the distances greater than 15m, air absorbs high frequencies [10]. However, this attenuation is relatively small - approximately 3dB to 4dB at 4kHz per 100m distance [52].

2.3 Spatial sound headphone reproduction

Spatial sound reproduction constists of technologies that enables the arbitrary placement of sound sources on a virtual auditory space. In principle the spatialized sound should have the same characteristics as a similarly located sound source in the real world. As in any reproduction system, the ultimate goal is to create a system, that is capable of producing virtual auditory scenes that the user would not be able to distinguish from the real world. A completely realistic, or *sonorealistic*, reproduction system is hard to implement due to the complex behavior of the real world sound fields. Yet, depending on the application, simplified systems that use approximations are often sufficient.

A typical headphone based spatial sound systems model the sound propagation to the two ears. These models are often used together with a room model [53].

2.3.1 Head-Related Transfer Functions (HRTF)

Head-Related Transfer Functions describe the propagation of sound in free field from the source to ear canal or to eardrum [11]. HRTFs can be measured or they can be constructed based on geometric models [26]. Measured HRTFs are referred to as HRTF-impulse responses or Head Related Impulse Response (HRIR).

HRTFs are individual. Measuring the individual HRIRs leads, in theory, to the best results and the localization accuracy can be equivalent or close to free field listening [54]. However, it is practically an impossible task to perform the HRIR measurements for each user of a spatial reproduction system. Hundreds or thousands of sampling points are needed in order to create a full HRIR set for one user. Furthermore, the measurements require special facilities and equipment.

2.4 HRTF individualization

There are several consequences of using non-individualized HRTFs. The virtual sound sources are not localized accurately and front-back confusions are often encountered [55, 56]. The elevation perception is particularly problematic [55, 57]. Furthermore, the sounds may not be externalized correctly and are perceived inside the head [25]. In general, the overall spatial perception may be distorted with generic HRTFs [58, 59].

The challenge is to find an efficient, practical and rapid method for individualization. As the HRTF measurements are expensive, it would be beneficial if the HRTFs could be obtained without them. If such a method existed, realistic 3D-audio and virtual auditory displays (VAD) would be available for a wide range of users [59].

2.4.1 Performance evaluation

A common subjective performance meter of a spatial reproduction system is a lateralization experiment [25, 58, 57, 60, 61, 62]. The lateralization experiment evaluates how accurately the subjects are able to tell the direction of a sound source. Other metrics that have been used are for example the externalization, front-back discrimination and the perceived realism [63, 62].

The performance evaluation is not a simple task. The goal is to present the sound object as realistically as possible - even up to an extent, where the reproduction system remains unnoticed and an illusion of being at the site is experienced. The lateralization experiments exclude this issue. It is often the case that when the lateralization is accurate, the spatializing is considered to be functioning. The overall quality of a spatial rendering system is a more complex issue than the plain lateralization accuracy. Rumsey has presented a scene-based paradigm to the subjective quality evaluation [64]. In this paradigm, the attributes are divided into micro and macro attributes. Micro attributes describe the features of single auditory elements, while macro attributes describe features of the whole auditory scene. These attributes are furthermore divided into four categories: width; distance and depth; spatial immersion, and a miscellanious category that includes attributes conserning source and scene stability and focus.

2.4.2 Measurement

The direct individualization method is to measure the HRIRs. Measurements often produces the best spatialization [63, 58, 56, 65, 66]. The outcome of the measurement is a database consisting of impulse responses that contain

information about the sound propagation to both ears at each sampled spherical coordinate. A specific HRIR can be accessed with the corresponding azimuth (θ) and elevation (ϕ) angles. The utilization of the measured HRIRs has a low computational cost and can the spatialization be performed in real time [61].

The measurements are performed under free-field conditions for example in an anechoic chamber [67, 68, 69, 11, 70, 71]. The subject is sitting or standing in the middle of a construction that is either a sphere, arc or a hoop [69, 72, 73]. The construction is a movable supporting framework for the speaker system. The speaker system produces excitation signal that is measured at both ears with miniature probe microphones [11, 72, 73]. The current standard measurement method is a blocked ear-canal or blocked meatus type, in which the microphone is placed at the entrance of the ear-canal [74, 58, 72]. The speaker system is moved and the measurement repeated until the full space around the subject is sampled [72]. The measurements are usually performed at a fixed distance.

The number of sampling points determines the spatial resolution. By using the suggested minimum audible angles (MAA) in the horizontal and vertical planes [75, 76], there will be approximately 2000 sampling points. As the localization accuracy is not evenly distributed, this number can be reduced by using non-uniform sampling points [72]. Furthermore, interpolation can be applied to intermediate positions, which can furthermore reduce the required number of sampling points [77, 78, 79]. The subject needs to remain at a static position during the process, which may take from two minutes [80, 81] up to hours to perform.

2.4.3 Database match

Instead of performing measurements for each individual, it might be reasonable to select the most suitable HRTFs from a database. For example, the CIPIC database, contains measurements from 45 subjects [72]. Other databases are the LISTEN database (IRCAM and AKG) [69], the FIU database (the Florida International University DSP Lab) [82] and the KEMAR database (MIT) [83].

The direct approach to database matching is to let the user choose the most preferable alternative from a selection of HRIRs [59, 63, 84]. This method evaluates indirectly the overall quality of the system. Other qualities of the reproduction system, such as externalization and elevation and front-back discrimination, may also be subjectically evaluated [63]. The individually measured HRIRs are often the most preferred, if they are included in the selection, but in some cases also the non-individual HRIRs are frequently

chosen [63].

Another approach is to use anthropometric data to find the best matching HRTFs [85, 86, 87] or to simulate them [88]. The anthropometric features may be automatically extracted with computer vision technologies [87, 81]. The CIPIC database contains 27 anthropometric measurement of head, pinna, neck and torso dimensions for each subject [72]. Furthermore, the HRTF database at FIU DSP Lab [82] contains 3D models of the pinna for each subject. For instance, the pinna parameters in the CIPIC HRTF database have been shown to have a significant effect on the lateralization accuracy [89, 90]. If the antropometric features are extracted from an arbitrary subject, they can be compared against a database. If a close match data between the extracted features and the features found in the database occurs, the corresponding HRTFs are, in theory, similar to the subject's individual HRTFs.

2.4.4 Anthropometric modeling

Approach, which aims to understand the influence of different body components on the HRTF is called structural analysis [27]. This approach is based on the proposition by Genuit in [91], in which each anatomical structure was represented as a filter. Due to linearity, the HRTF can be constructed as a cascaded combination of the filters [27, 92]. Structural modeling is attractive, as it enables computationally inexpensive real-time HRTF rendering [93].

The pinna is amongst the most influential body parts in the spatial hearing [10]. The effect of the pinna on the frequency response and localization has been widely studied and modeled [94, 95, 27, 96, 11]. Pinna reflections produce several notches into the spectrum, that can be seen in the pinna related transfer function (PRTF) [93, 92, 96]. The PRTFs may be measured, modeled by using anthropometric data or extracted from HRTF data.

The simplest structural HRTF model is the head and torso model (HAT) [97]. It is constructed by cascading the PRTFs with simple spherical models for head and torso [98]. HAT is an accurate approximation of the HRTF especially below 3kHz and it is also capable of providing cues for elevation perception [99].

Chapter 3

Auditory interfaces

The main senses that the human uses to obtain information about the world are vision and hearing. Vision is the dominant modality and the current culture is highly visual. Hearing is easily neglected as in the everyday listening it remains somewhat *hidden* in the background. The sound can go unnoticed when it is present, but when there is no sound, we notice it immediately and we might feel even disoriented.

A large amount of information is conveyed via sound in our daily activies. The hum of an air condition device, approaching foot steps and almost an unlimited number of other examples. We are able to learn the finest details of sounds, and we are therefore, for example, able to tell the approaching person by the foot steps, and if there something wrong with the air conditioning device. What is remarkable is that these observation processes are mostly autonomous [20, 11]. Indeed, the sound as a modality is different to visual. We do not need to turn our head to focus our auditory attention on something - this is a cognitive ability. Furthermore, sound is of spatiotemporal nature. It seldom remains constant in time.

Auditory interfaces use sound to present information [100]. The research on auditory interfaces attempts to find new methods and areas in which the use of sound is beneficial. The research incorporates knowledge of the properties of hearing and sound, cognitive issues, interaction methods and sound reproduction technologies.

Auditory displays are not a new subject. The earliest example is from the mid 1800 - the Morse code [101]. Trained people are able to decode the message in real-time without no real effort. Furthermore, sound is used in many graphical user interfaces (GUI) to provide feedback. The research of auditory interfaces has been gaining and increasing attention during the last two decades. The International Community for Auditory Display (ICAD) was founded in 1993.

3.1 Usage of sound in interfaces

Sound may be used as the primary or as a complementary modality [100]. The latter case is more common. For example, sound is often used in conjunction with vision in the GUIs, in which sound may for example reinforce the visual information. The design of an audio-only interface follows very different paradigms as the nature of sound is very different to visual. It must be considered why to incorporate sound and what kind of information is presented via sound [100].

Using sound in interfaces can provide many benefits. It can reduce the visual overload caused by many GUIs [100]. Sound can also efficiently draw the attention from one task to another [100, 11]. The auditory system is sensitive to detect patterns in sound. Thus, sound is powerful in exploratory data analysis of large data amounts e.g. seismic or medical data. Sound can furthermore convey emotions.

Sound should be used with care. If the sound is poorly designed, uninformative, non-relevant, used too frequently or has bad quality the user may get annoyed and even turn the sound off [100]. The aural modality may even be more sensitive to overload than the visual. Also, the amount of concurrent sound events should be carefully considered, as it quickly increases the masking and cognitive load [102, 100]. In the worst case, the sound is interpreted as noise [100, 11]. The interface should benefit from using sound.

3.1.1 Sound as a complementary display modality

Sound events in computer interfaces are common. There are distinct sounds that are played when e.g. an error occurs or when the user logs in or shuts down the computer. These sounds are typically action indicators. Usually the user is able to change the sound theme or the sounds. An early example of using sound a complementary modality is the SonicFinder by W. Gaver [103], which was created in 1989 for Apple. The SonicFinder used auditory icons to present events and actions. Gaver's approach was very fundamental and the legacy of his work can be seen on the current operation systems Apple OS X and Windows [104]. Furthermore, sounds are used as a complementary modality in mobile phones, healthcare systems, radars and in an increasing variert of other devices and interfaces [100].

Sound is used for many different purposes. Among these are providing additional information when vision is occupied, managing user attention, providing feedback, process monitoring and sensory substitution [100]. Sound is especially efficient for attention managing. A sudden sound or change in

the sound immediately draws the attention [11, 105]. This property applies for process monitoring and warning systems. An example of a monitoring system is the ECG (electrocardiogram) heart monitor. The ECG produces a repeating sound of the heartbeat often along with a visual representation. The user adjusts easily to the sound, but any innormalities are instantly detected.

Humans integrate information from multiple modalities [106]. As a task or the information is divided upon multiple modalities, the cognitive load is reduced. For example it is easier to understand what someone is saying if we see his lip movement [107]. As the complexicity of a task increases, the more the multimodal actions take place [108, 109]. Furthermore, as the same information is conveyed via two modalities the interface becomes more effective [100, 110]. Modalities can also substitute each other [100, 111]. Vibrotactile feedback may be substituted for example with sound and vice versa.

3.1.2 Sound as the primary display modality

The majority of products that use sound as the primary modality are for the visually impaired people [100]. These products are aimed to aid the everyday actions and include screen readers, navigation systems and accessibility for desktop and mobile computing [112]. Screen readers present the contents of a screen as sound. The sound may be speech or any other sound. In the modern GUIs, the mapping of the screen contents to sound is not a straightforward process. One solution is to add a low-level layer called off-screen model (OSM). System messages, visible and hidden screen contents are stored into OSM, analyzed and then read aloud to the user [113].

Accessibility has been considered in the major operation systems. Microsoft Windows has included the Microsoft Narrator screen reader since Windows 2000. Apple has developed the VoiceOver screen reader, which was first introduced for Apple desktop computers in 2005. VoiceOver is currently available also devices using the iOS and to the small iPods nanos. Though, the most popular screen readers i.e. JAWS, Window-Eyes, Dolphin, are separate commercial products [114, 115, 116, 117].

Navigation is another feasible area for sound [118]. Besides, being beneficial for the visually impaired, it would also benefit the normally sighted. Wayfinding consists of sensing obstacles and hazards in the environment and navigation. As we consider the case where a person is in a new location with a navigation task, there are actually two tasks that require vision: scanning the immediate environment and occasionally reading the map. As the person looks at the map, disorientation and even hazardous situations may occur

[119]. If the correct direction is aurally presented, the vision would remain on the immediate environment. Sound based navigation systems for the blind have been studied e.g. in [120].

A completely different product category are PMDs that do not contain a screen or contains a small screen. As a disadvantage of being small, the devices often suffer from diminished usability [100]. These PMDs would greatly benefit from an auditory interface. There are auditory interfaces designed for these purposes. Two examples are the earPod [121] and the Funkyplayer [122]. Both are touch-controlled and incorporate the use of spatial auditory menus, where the information and the auditory objects are auralized and spread around the user. This type of spatial distribution is referred to as radial pie.

Alarm systems are another classical example of systems that uses only sound [123]. Fire alarm and air alert systems draw efficiently the attention and carry the message even to distant locations.

3.2 Perceptual dimensions of sound

Sound is a physical phenomenom. The auditory system has evolved to detect even slightest variations in sound. Psychoacoustics is a research area that studies these sensations caused by sound [11]. The auditory system maps the variations in sound into several psychoacoustic quantities e.g. the perception of loudness or pitch. Some of the psychoacoustic quantities are seemingly simple, while others may be more complex and can vary due to cultural background, e.g. the sensory pleasantness [11]. Even more complex issues are involved in the auditory scene analysis, which attempts to describe how the auditory system separates single auditory streams in an auditory environment [11, 124].

3.2.1 Psychoacoustic quantities

Pitch

Pitch describes the perceived sound frequency [11]. It is primarily depended on the sound frequency, but also other factors affect e.g. the complexity of the tone [11, 20]. The hearing range for a healthy person is approximately from 16Hz to 20kHz. The resolution is frequency depended. Below 500Hz the JND for a sine tone is 3Hz and 1Hz for complex tones. Above 1kHz the JND for sine tones is approximately 0.6%. For example at 2kHz the JND is approximately 12Hz [125].

Loudness

Loudness is the perceptual equivalent to sound intensity [126]. It has been found that the sensation of sound intensity is also frequency dependent [11, 20]. In the case of sinusoidal tones, this means that sounds with different intensity at different frequency bands cause similar loudness sensation. This is visualized in Figure 3.1, that contains an so called equal-loudness contours. The loudness sensation of a counter is equal along the frequency axis. The auditory system is most sensitive at speech frequencies ranging approximately from 100Hz to 7000Hz [11, 126].

From an auditory interface viewpoint, loudness variations, i.e. dynamics, could be used for example as a parameter for information urgency [100].

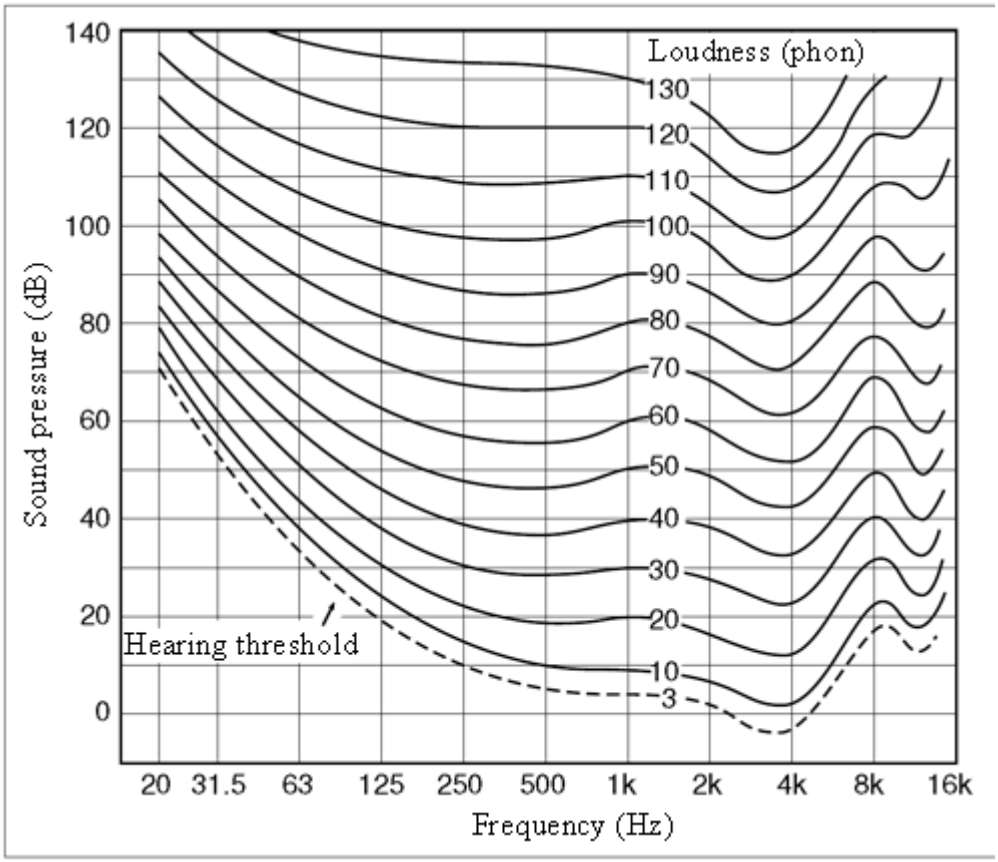

Figure 3.1: Equal loudness contours. Adapted from [11]

Timbre

Timbre describes the perceived spectral properties of a sound. Unlike loudness and pitch, it is a multidimensional quality that cannot be presented as one number [11]. Timbre is the quality that enables the auditory system to distinguish between two instruments that play the same note. The sensation of timbre is of a complex spectrotemporal nature. For example the envelope, harmonics and modulation all have an impact on the timbre [11, 20]. Especially the attack time is a crucial factor in distinguishing wind instruments [11]. In auditory interfaces, timbre can be used for efficient sound source separation.

Spatial properties

All real world sound sources have spatial properties. The auditory system is able to determine the direction, elevation and distance of a sound source [10, 11, 20]. Also the space in which the sound event occurr is integrated into the sensation [11, 12]. As there are multiple concurrent auditory streams, the auditory system is able to more efficiently focus on one stream, if the streams are spatially separated [127, 100, 128]. In particular, spatial separation of concurrent speech sources increases the speech intelligibility [129]. Furthermore, spatial properties adds a new dimension to the sound. As the sound and the corresponding location are grouped, the memorizing of an user interface (UI) element may be improved [130]. Thus, adding the spatial dimension may also improve the creation of a cognitive spatial map for the interface [131]. Also, the interaction becomes richer, as there are more dimensions.

Masking

Masking is an aural counterpart to occlusion. In vision, if a larger object is in front of a smaller one, the smaller becomes invisible. But if a loud low frequency tone and a soft high frequency tone are played at the same time, the louder tone does not necessarily mask the softer one. If the tones are close to each other in frequency, masking is likely to occur. In a physiological sense, a tone with a certain frequency activates the sensory cells at a certain spot on the basilar membrane [11, 132]. This is the so called *place theory*. If two tones are aligned closely to each other on the basilar membrane, the capabilities of the sensory cells to detect both is diminished. The tone with higher energy might completely override the weaker one [11]. The term *critical band* describes this type of frequency resolution of hearing. Both loudness and frequency affect masking [11].

Masking has also a temporal dimension. It affects to sounds 5-10ms before and 150-200ms after the masking sound [11].

Tonality, consonance, dissonance and distortion

Tonality means two closely related things. In psychoacoustics, a tonal sound contains narrow bandwidth components from which the fundamental frequency or a single voiced parts can be detected [11]. In a musical context, tonality refers to the concept of having a fixed tuning system [133].

Sensoric consonance and dissonance are quantities that describe as how pleasant or unpleasant two concurrent tonal sounds are perceived and how well they play together [11]. Consonance and dissonance are determined by relationship of pure tones of the two sounds in relation to the critical band [11, 20]. In auditory interfaces, dissonance could be used to express that something is wrong or out of place.

Distortion is, generally, any perceived irregularity of a sound [11]. It requires a reference. Distortion is not directly related to consonance and dissonance, but it can have similar metaphors in auditory interfaces.

Tonality at large, can be used for many purposes in auditory interfaces. For example, the relation between consecutive notes is easily detected and melodies are effectively memorized. Furthermore, music and certain sounds have the ability of conveying emotions [100]. Emotions can create strong, lasting memories [134, 135]. If an emotion is associated to a part of an UI, the learning process could be more rapid and efficient.

3.2.2 Auditory scene analysis

Among the most remarkable properties of the auditory system is the capability to analyze complex auditory scenes. Even if the auditory scene contains multiple concurrent auditory events, the overall scene is not perceived as confusing [10]. Rather, the auditory scene is comprehended as a whole - each sound has its own place. We can focus essentially at one sound source at a time, but we can also rapidly change the focus point between many sources [11]. This is the so-called *coctail-party effect* [136]. Furthermore, by using the auditory short-term memory we can separate and track up to three simultaneous sources [11]. This is often the case when listening to music.

The capability of the auditory system to analyze the aural environment is partly in-built and partly learned [11]. We have for example learned that an approaching car produces a certain set of sounds. These sounds are associated into the same object, which is then perceived as a single sound source.

As the sensory input, including visual, aural, haptic etc., is enormous, filtering is required as the processing capacity is limited. The cognition automatically filters and organizes the sensory streams and is able to locate new, relevant and unfamiliar details [137]. The majority of familiar information

may remain unregistered, thus reducing the cognitive load. The cognition groups different events according to some properties. These issues are studied in cognitive, experimental and gestalt psychology. The gestalt psychology explains eight of objects grouping principles in the sensory domain [11, 138]. Five of these principles are

- Principle of proximity. Objects that are close to each other space are likely to be grouped as to the same object.
- Principle of continuity. Objects close to each other in time are likely to be interpreted as part of the same causal continuum. Unless opposite evidence is presented.
- Principle of similarity. Similar objects are grouped. Irregular objects are easily observed from a pattern.
- Principle of closure. Objects may be perceived as whole, even there might be missing parts.
- Principle of common fate. Although having a dramatic name, it refers to the trajectories of moving objects. Objects and elements that have same trend of motion are likely to be grouped.

3.2.3 Cognitive load

The cognitive load is a psychological term that refers to the executive control of the working memory. The cognitive load theory (CLT) states that the working memory and the number of cognitive operations that can be performed on it, is limited [139, 140, 141]. As there are multiple simultaneous sound sources and tasks, the cognitive load rises. The load can be measured with the NASA task load index (NASA-TLX) [142]. The CLT theory futhermore suggests that the limitations of the working memory can be circumvented up to some extend by coding the information as a one element in a cognitive schemata, that is a mental structure of a part of the real world [143]. In the context of auditory interfaces, this would yield a careful consideration of data structuring and presentation.

3.3 Mapping information to sound

The essence of auditory interfaces and displays is to efficiently map information to sound. There are several technologies and approaches. For example sonification is a set of methods that attempts to create an audible, even

multidimensional, representations of a given set of data or an interface [123]. The output of sonification is often abstract. Nevertheless, non-abstract sound objects are also frequently used. These sounds may contain speech or they can be recognizable in a sense that they resemble real-life sounds.

Before applying any sonification or synthesis technologies, two fundamental question need to be solved. *What information should be presented*? As the human auditory and cognitive capacity is limited, this question should be carefully considered. Another critical issue is *how the information should be presented*? As a comparison, the GUI desktop is developed on the metaphor of an actual desk. There is the recycling bin, folders and icons representing several other objects that can be find around a desk. Clearly, this metaphor has no direct use in the context of auditory interfaces. What kind of metaphors should be chosen? Although the research field is relatively new, there are no standard guidelines for mapping strategies or metaphors in auditory displays [144]. It might be due this, that there has not been a major auditory interface breakthrough. As a result, only few examples of simple consumer auditory interfaces exists, majority being scientific curiosities.

3.3.1 Sonification

Sonification is the use of sound to convey information [100, 123]. It is a common method in auditory displays. Systems that use sonification are for example the electrocardiogram (ECG) machines in hospitals, the Geiger counter and proximity radars that are used in cars [145]. Trained medical staff are able to predict and therefore avoid seizures by monitoring the auditory output of the ECG machines [146]. Further examples include the mapping of stock market price [147, 148] to sound and monitoring computer network services. Depending on how the the sonification is implemented, the result may have for example musical qualities (rhythm, timbre, melody), psychoacoustic qualities (pitch, loudness, spatial location, dissonance/consonance) and any combination of these. A major benefit of using sound is that patterns and irregularieties are efficiently detected [123, 100].

Sonification as a concept was introduced at the first ICAD (international conference of auditory display) conference in 1992 [145]. The early definition stated that

> ***Sonification*** *is the use of non-speech audio to convey information.*

The definition was problematic as it did not provide a systematic nor scientific ground for sonification [149]. Furthermore, the speech was excluded.

Later, a more detailed definition was given. The definition includes other sound objects and concepts e.g. auditory icons, earcons and audification for they can be used for sonification. Furthermore, music is excluded. The new definition as presented in [149] states:

> *A technique that uses data as input, and generates sound signals (eventually in response to optional additional excitation or triggering) may be called* ***sonification****, if and only if*
>
> 1. *The sound reflects objective properties or relations in the input data.*
> 2. *The transformation is systematic. This means that there is a precise definition provided of how the data (and optional interactions) cause the sound to change.*
> 3. *The sonification is reproducible: given the same data and identical interactions (or triggers) the resulting sound has to be structurally identical.*
> 4. *The system can intentionally be used with different data, and also be used in repetition with the same data.*

The main sonification technologies are audification, parameter mapping sonification (PMS), the model-based sonification (MBS) and symbolic sonification [123, 100]. In audification, each data value is used as a sound signal value [145]. The method essentially requires a large amount of data [100]. An example of auditory displays that use audification is the ECG machine.

PMS is conceptually close to the visual scatter plot, where the features of the data determine the graphical output, e.g. the x- and y-position, size and color [145]. The input and output have a direct correlation. In PMS the features can be mapped into different psycoacoustic parameters, thus presenting multidimensional data. By using the temporal dimension, large data-sets can be efficiently represented. This can be referred to as the exploratory data analysis [123].

Model-based sonification creates a dynamic model into which the data is integrated [123]. The metaphor that is used in MBS is very different from PMS. As in real life, objects may be constructed from several materials and they can contain various types of structures. If we hit a drum, the resulting sound depends on the drum materials, dimensions, how hard the drum is hit and to where it is hit. MBS follows this metaphor and creates dynamic sound objects that are excited by the user [100, 123]. The data is an internal part of a model. The *Shoogle* [150] is an interface for mobile devices that uses

MBS for sensing data within the device. As the user shakes the box, different emails (read, unread, long and short) generate a different type of sound as they hit the wall of the box. The data itself becomes an *instrument* [123].

3.3.2 Symbolic sonification

Symbolic sonification is different by nature from the audification, PMS and MBS. PMS and MBS produce a dynamic output, whereas the symbolic sonification use single sounds [151, 103]. The symbolic sounds can be abstract or non-abstract. Some of these concepts e.g. the auditory icons are older than the first definition of sonification [145, 151].

Auditory icons

Auditory icons attempts to create an intuitive metaphorical linkage between the objects and events in the interface objects and the real world sounds [123]. The method should enable a rapid and effortless identification of the instance [100]. Thus, the learning curve should be narrow. Furthermore, the interface objects and events do not always have a real world equivalent sound. In these cases, other iconic representations can be considered. The auditory icons can be recorded or synthetic sounds [100]. Furthermore, the auditory icons can be parametric [123]. Small bouncing ball may represent a short message, whereas a large ball a long message.

Earcons

Earcons are short, structured musical messages [145, 118]. An earcon may be a series of notes. The pitch of consecutive notes can be rising or descending. Different musical properties, such as timbre, rhythm, loudness and pitch, are be associated with different properties of the data or event. The main difference between earcons and auditory icons is that the earcons do not imply a relation between the sound and represented information. Thus, the earcons need to be learned. As a benefit, structured earcons are a type of highly abstract symbolic form of communication [100].

3.3.3 Speech

Speech is frequently used method to present information in auditory displays and interfaces. It has been used for long in screen readers and computing for visually impaired in general. Speech is very efficient in presenting quantitative, complex and precise information. Indeed, in many cases it is impossible

to present the information ambiguosly in symbolic or musical terms. As a disadvantage, speech is a slow method and it may become tiring and irritating, especially if the message is not relevant [152]. Furthermore, speech does not provide possibilities for abstraction. As a final note, the information itself, does not always have to be excact or absolute.

Speech can be produced either by using recorded or synthetic speech. In most of the cases, recordings are more natural and pleasant, but the workload and memory requirements increase rapidly as a function of the speech database complexity. Furthermore, recorded speech is not dynamic which makes it obsolete in many cases.

Speech synthesis is a more versatile option. Systems that are able to create artificial speech out of written text are referred to as text-to-speech (TTS) systems. The TTS procedure consists of two main phases. These phases are a high-level analysis of the text and speech synthesis [11, 153]. The high-level analysis is performed in a natural language processing module [154]. As an outcome of the analysis, the text is transcribed into a phonetic representation, that can contain additional information i.e. prosody, intonation and stress [153, 155] that are normally present in natural speech.

The speech synthesis methods are usually divided upon three categories. These are the concative synthesis, formant synthesis and articulatory synthesis. Concative synthesis is the most simple method, but it also provides the most natural speech [20, 156]. It uses a prerecorded speech corpus. The corpus may contain material from single phonemes and words to sentences. The output is then generated by choosing the corresponding speech segments in order to create a phonetical representation of a written sentence. Concative synthesis has a limited vocabulary.

Formant synthesis is a commonly used, efficient and unrestricted method. It is based on a source-filter model of the speech [11, 153]. Formants are the resonant frequencies of the human vocal tract. Three to five formants are generally needed in order to produce intelligible vowel-sounds [157]. The two main resonator structures are the cascade and parallel [11]. Parallel structure is more versatile, as the amplitude and bandwidth of each formant can be separately controlled [11]. The excitation signal is periodic glottis pulse for voiced and noise for unvoiced sounds [158, 11].

Spearcons

Spearcon stands for speech-based earcon. A spearcon is a word or sentence that is sped up, without changing the pitch, so that it is no longer comprehensible as speech [159]. Spearcons are beneficial over earcons as they can be produced using TTS software. Furthermore, it has been shown that the learning process

using spearcons is faster than of earcons [160, 161, 162]. In fact, they are as easy to learn as speech [163]. This is due to the fact, that spearcons have a non-arbitrary correspondence to the item that they are representing.

3.4 Auditory menus

Auditory menus present hierarchical structures using different sonification methods. Unlike visual menus, auditory menus do not have a standardized design methods or guidelines [164]. The most common implementation is a speaking-menu, in which TTS is usually used to present the different items [164]. Other methods include auditory icons, earcons and spearcons, but these are generally used for research purposes [164]. Different methods can be complementary to each other. For example earcons can be used to provide contextual information and feedback. Different menu levels may be presented by playing a certain sounds or note sequences.

Generally, one item is presented at a time. The number of identifiable and discriminable sounds decrease as the number of concurrent sounds increase [165, 166, 167]. However, the discrimination can be improved, if the sounds do not have the same onset times [168] and they are parsed upon multiple auditory streams in the auditory system [169, 124]. For example if the onsets of two concurrent sounds, timbral characteristics or spatial positions are different, the sounds are likely to parsed upon different streams [168, 124, 11]. Thus, spatial sound has been used in auditory menus. For example, NASA has been investigating the use of spatial sound in human interfaces for almost three decades [170] The main spatial mappings are horizontal and vertical [121, 122, 171]. In the horizontal mapping, the menu items are placed on a ring around the user. In the vertical mapping, the elevation is altered [167].

3.5 Ambient auditory displays

Ambient auditory displays convey information by using the metaphor of ambient soundscape. The information is sonified and displayed as a non-distractive and non-intrusive part of the natural soundscape. This metaphor is particularly interesting, as it is closer to a real life listening. We are able to build an efficient mental image of the environment by passively listening to the background sounds, even while we are concentrating on our primary tasks. Furthermore, the auditory system adjusts to the soundscape, but any interesting irregularities are able to draw our attention [172].

Ambient auditory displays are particularly interesting in the context of

this thesis. The amount and of information that the PMDs are able to receive and collect is vast. Ambient auditory displays may provide a non-instrusive method to display that information.

3.5.1 Definitions

Ambient auditory displays have common characteristics with the peripheral systems and notification systems, and are a sub-category of ambient information systems [173, 174, 175, 176, 177]. As defined in [178], the ambient displays have the capability to

> *Ambient displays present information within a space through subtle changes in light, sound or movement, which can be processed in the background of awareness.*

Awareness is another particularly interesting term in the context of auditory interfaces. It is a state of knowing about the surroundings and activities that takes place in it [178].

There are five behavioral characteristics for ambient displays [177]. First is their capability of displaying information that is important but not critical, which differentiates them from alerting displays. This also makes them ideal for tracking and being aware of background processes, which is, in other words, multitasking. Second main characteristic is that the user can move his focus of attention back and forth from the ambient display. Third characteristic is that ambient displays should be not distracting. Information is conveyed through intuition rather than interruption. The ambient display might be continuously be present, thus fourth property is that the ambient display should be aesthetically pleasing. Last of the five properties is that the ambient display should blend into the environment.

3.5.2 Implementations

First ambient auditory display was introduced by Jonathan Cohen in an application called ShareMon [179], which used sound to notify the user of events concerning file sharing. Cohen wanted to be aware of *"what is going on behind my back"*, on a computer UI. He came up using a foreground and background tasks theme. Foreground tasks are under active consious control. The background processes execute automatically without user intervention, but system reminds the user of the state of the processes. The foreground and background activities are discussed in [130] from a cognitive psychology viewpoint.

A more recent attempt to create a non-intrusive example display is the Weakly Intrusive Ambient Soundscape for Intuitive State Perception (*WISP*) [180, 181]. *WISP* creates a forest soundscape, consisting of bird calls, into a personal or shared space. A new bird appears into the soundscape when a co-worker arrives and disappears vice-versa. The design guideline was that the user should not be distracted by the presence of the soundscape. WISP is intended to be a part of a larger setting, where a whole variety of computer software and facilities for graphic and audio presentations interact with the user in a coordinated fashion. A similar, soundscape based work was presented in [182], which studied awareness and lightweight interactions.

Chapter 4

Gesture controlled auditory menu

This chapter present an eyes-free interaction method that was designed to control an auditory menu. The tactile gestures are detected by using acoustic recognition. The gestures are detected through the fabric so that the device is accessible at all times. Furthermore, a traffic simulation test is presented that can be used to compare different types of interfaces.

Many of the tasks that the users perform on a PMD are relatively simple. The user might be for example changing a track on a music player or checking tomorrows weather on a phone. The needed tasks for these actions require taking the device out of the pocket (access time) and performing the actual task (usage time) [183]. Usually, the access time is longer than the usage time for these simple tasks.

In order to shorten the overall interaction time, there has been plenty of research on new interaction technologies and paradigms. Also, new terms have been coined: microinteractions [183] and always-available interaction (or input) [184]. For example, Nokia has presented "Tap input", which is a minimalistic interaction method for mobile phones [185]. A 3-D accelerometer was used to measure the direction of the taps, shakes and knocks that were used in simple interaction tasks. WhackGestures [186] presented a similar idea of using an accelerometer to detect "whacks" and "wiggles" for microinteraction with the emphasis of minimizing the device access time. PocketTouch [187] used capacitive sensing to detect multitouch finger input through the fabric.

While these were for quite harsh and inexact interaction, there are examples of work that provide more accurate interaction. Fingerpad [188] transformed the index finger into a track-pad using a 3x3 grid of Hall-sensors and a magnet that was placed to the thumb. Another fingertip based approach combined a small LED and a camera into a device named MagicFinger [189]. MagicFinger was able to detect gestures (tapping and position) and different surface textures. Artificial surface textures were intended to be used

for different interaction contexts - for example for sliders and input stickers. Stane [190] was a prototype of a hand-held microinteraction device that used acoustic sound recognition to detect tapping and scratching of a rigid surface. A piezomicrophone was inserted inside the device to detect vibrations that were caused by touching. The device contained a textured rotary wheels with varying frequencies and gradients.

This structure of this chapter is as follows. First, the prototype and the sound recognition module are presented. After this, an auditory menu design along with a gesture to mapping sutudy is introduced. Finally, the performance of the gesture controller method is compared against a visual menu in a traffic simulation test.

4.1 The gesture controller

The focus of the design was to enable eyes-free always available microinteraction for completing simple tasks on a PMD without hardware modifications or external controllers. Furthermore, current work emphasises the use of sound as the primary feedback channel.

The controller was designed to detect and classify the acoustic features of four tactile gestures. In general, four commands are needed to navigate in a hierarchical menu (next item, previous item, select, cancel). The gestures were *tap*, *double tap*, *swipe* and *double swipe*. These gestures were selected as they are rapid and easy to perform and they are also used in smartphones.

4.1.1 The physical controller

The prototype is a piece of dense plastic foam (3x6x1.5cm). The piece was cut by hand. A hole was carved into the foam. A wireless lavalier microphone (Sennheiser EW100) was inserted inside this hole. The surfaces of the piece were thickened with hard paper. The wireless microphone was used to transmit the contact sounds generated by the four gestures to a computer. The finalized prototype is shown in Figure 4.1. The controller resembles a PMD in size so it can be placed inside a pocket.

4.1.2 Sound analysis

As the physical controller was constructed, the sound produced by the four gestures (tap, double tap, swipe, double swipe) were recorded and analyzed. Seven subjects were invited to perform the gestures three times while sitting, standing and walking. The controller was placed inside the pocket of a subject.

Figure 4.1: The protoype gesture controller.

The tap was performed very uniformly, there were some variations on the swipe gesture.

4.1.3 Temporal characteristics

Examples of the recorded sounds are presented in Figure 4.2 and Figure 4.3. Several temporal stages can be identified in the figures. These stages are marked as A, B, C, D and E, and they are described in Table 4.1. The average durations of these stages for each subject are presented in Table 4.2.

The tap gesture produces an impulse type sound that attenuates rapidly. The average durations of the taps (part A) varied from 31ms to 45ms. The duration is slightly depended on how sharply the gesture is performed. For each subject, the average tap duration was slightly longer while performing the double tap. Delay between the consecutive taps (part B) in the double

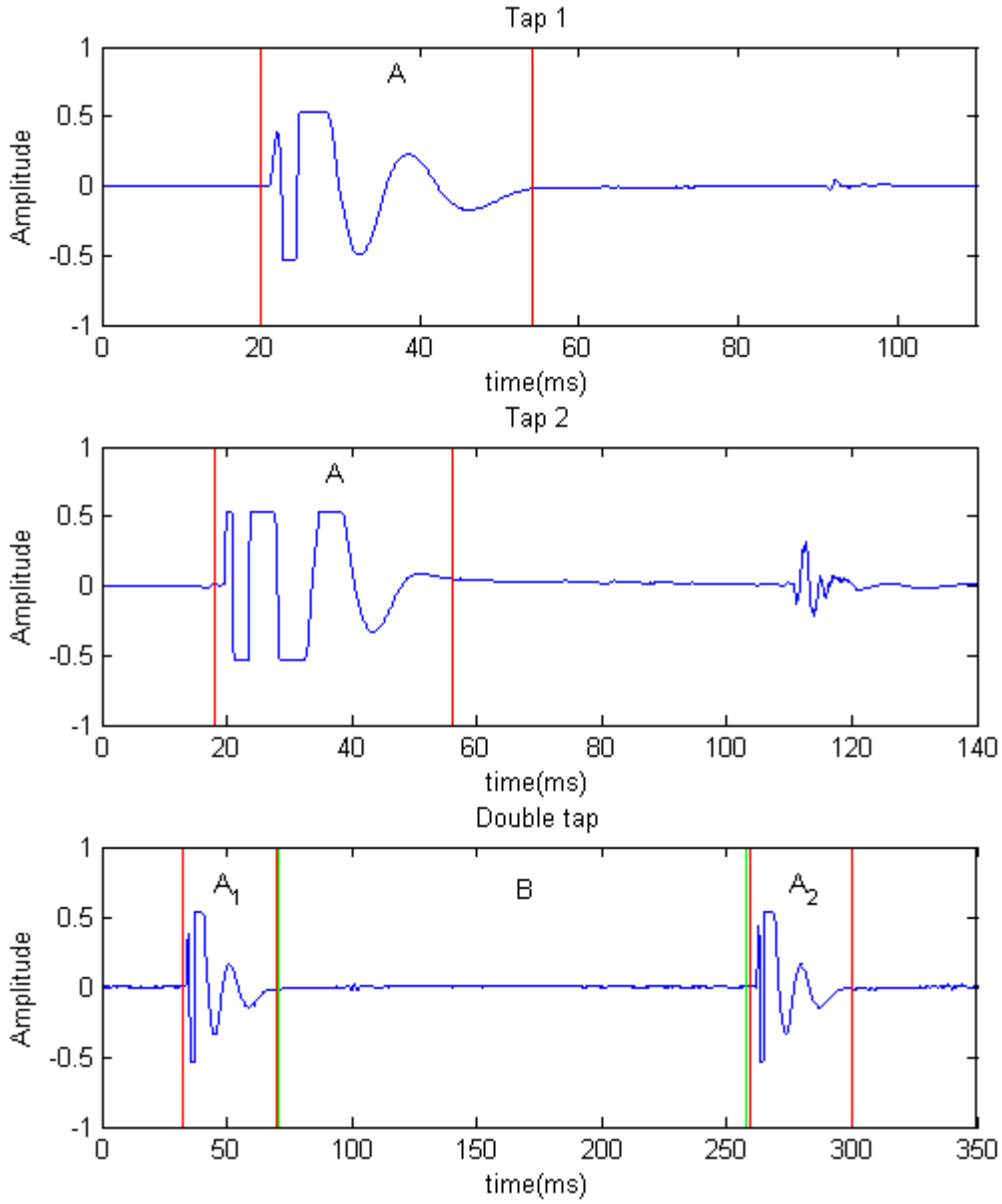

Figure 4.2: Waveform of the tap and double tap gestures.

tap varied from 145ms to 267ms. Some subjects performed a tap so that another, softer, impulse appeared approximately 80ms after the first impulse. An example of this second impulse can be seen in the subfigure 'Tap 2' in Figure 4.2.

The swipe generates a noise type signal, that has a significantly longer

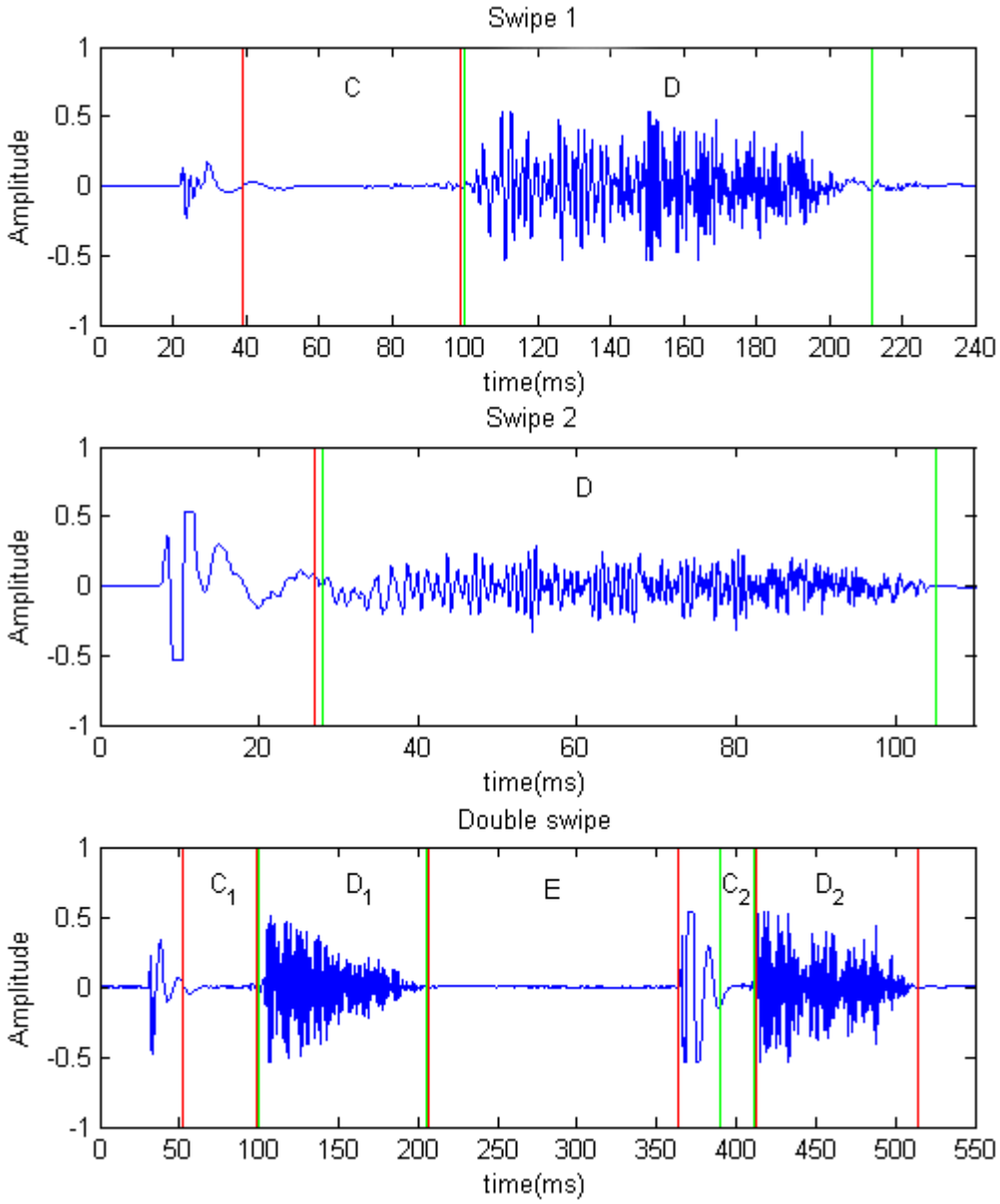

Figure 4.3: Waveform of the swipe and double swipe gesture.

duration than than the tap. The average duration of the gesture (part D) was approximately 170ms. The duration of a swipe was shorter in the double swipe. Some subjects performed a soft *contact tap* before the swipe gesture. The duration between the contact tap and the actual gesture is marked as part C on the subfigure 'Swipe 1' in Figure 4.3. Average length of the part C

varied from 12.6ms to 119ms and it appeared more frequently in the double swipe gesture. The delay between two consecutive swipes (part E) was of the same magnitude as in the tapping gesture.

Part	**Description**	**Gesture**
A	Impulse duration	Tap
B	Delay between the two taps	Tap
C	Contact delay	Swipe
D	Swipe duration	Swipe
E	Delay between the two swipes	Swipe

Table 4.1: Temporal stage descriptions of the gestures.

Part	**Subject**						
	1	2	3	4	5	6	7
A (single)	38	33.7	31	38	39.7	32.3	35.7
A (double)	40.8	34.7	40.8	45	43.2	37.5	38.2
B	187.7	145.3	263.7	151.3	199	199.3	248.3
C (single)	119.3	0	0	0	44	12.6	0
C (double)	61	19.3	4.3	0	39	27.5	0
D (single)	179.7	346.7	-	91	123	257.3	43.3
D (double)	149.8	211.5	172.8	64.5	107.8	226.3	42.3
E	216.3	176.7	161.3	177.3	154	232.3	238

Table 4.2: Average durations (ms) for the gesture temporal stages.

4.1.4 Spectral characteristics

The frequency content for tapping and swiping were distinct. The spectra of both gestures are presented in Figure 4.4, where the top subfigure presents a tap and the bottom subfigure presents a swipe. The tap generates a spectrum that ranges approximately from 60Hz to 400Hz. The spectrum had a peak at approximately 80Hz.

Swipe has spectrally more broad energy content. Most of the energy is concentrated at frequencies ranging from 80Hz to 1200Hz. Furthermore, four peaks, located at 40Hz, 80Hz, 400Hz and 900Hz, can be seen in the bottom subfigure of Figure 4.4.

As the controller was designed to be held inside a pocket, movement generates noise due to the rubbing of the clothes. This noise has the majority

of its energy at frequencies ranging from 30Hz to 300Hz but it also has smaller components up to 2000Hz. The energy content of this noise is at the same frequency domain as tapping, but the magnitude is significantly lower.

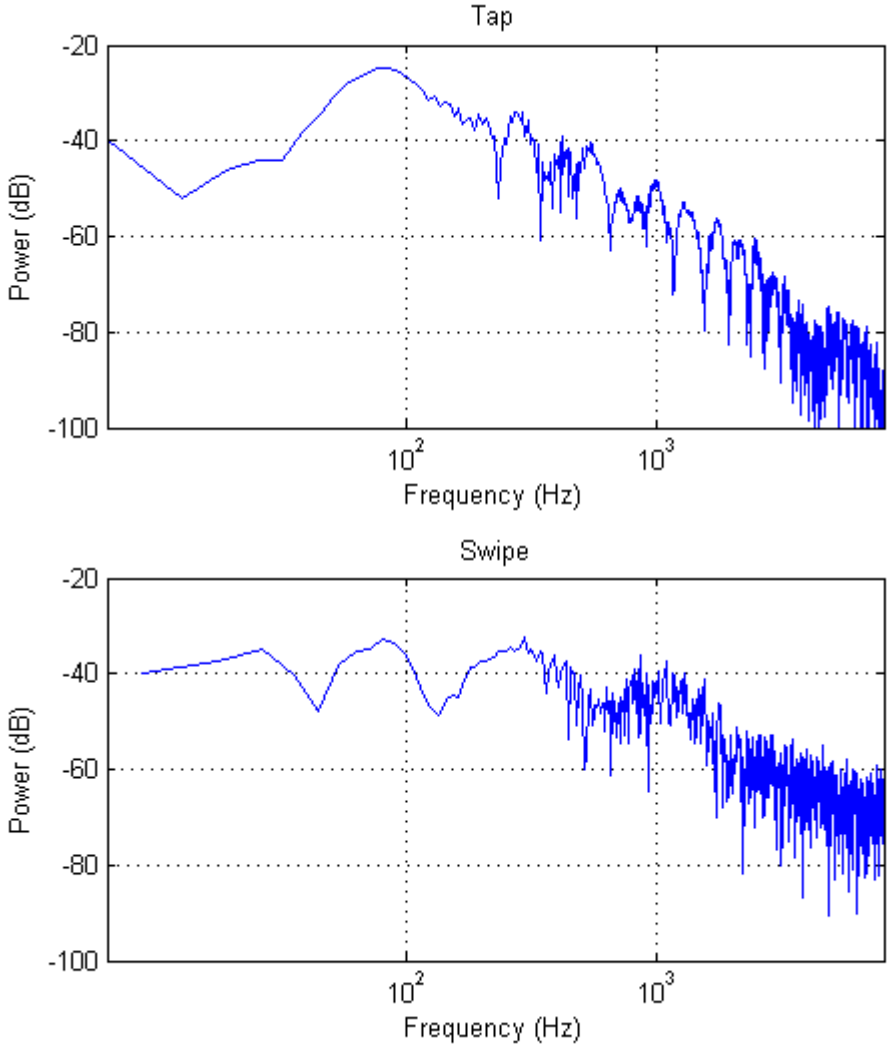

Figure 4.4: Spectrums of the tap and swipe gestures.

4.2 Acoustic classification module

The acoustic classification is based on the duration and spectral characteristics of the gestures. In general, tap produces a short impulse type sound that has a the majority if energy below 400Hz. Swipe produces a significantly longer signal that has more energy at higher frequencies.

The system consists of a filter block, gesture duration and root mean square (RMS) power measurement, noise gates, timer and decision logic. Microphone signal is first fed into the filter block, where it is divided into two frequency bands. The average RMS power of 60ms segments were measured at both bands. The RMS levels were set to open and close the noise gates. The time that the gates remained open was measured. The final classification was performed at the decision logic. The block diagram of the system is shown in Figure 4.5. The real-time system was built in the Max/MSP 5.0 programming environment.

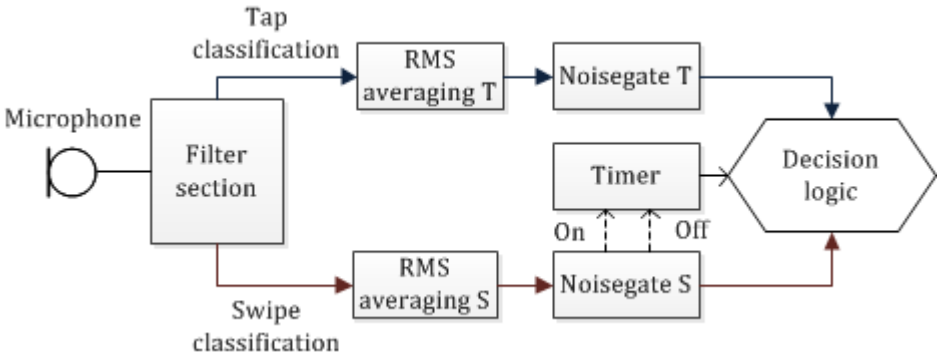

Figure 4.5: Gesture classification module block diagram.

4.2.1 Filter block

The filter block divided the microphone signal into a tap and a swipe classification bands. These bands were derived from the analysis shown in subsection 4.1.4 and were iteratively fine-tuned. The block was built using biquad filters. Biquad filter is a common name for a two-pole, two-zero filter. The transfer function of a biquad filter [191] is

$$H(z) = \frac{1 + b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}} \tag{4.1}$$

Since the energy distribution of the tap gesture is concentrated at low frequencies, the signal was first low pass filtered using a cutoff frequency of 110Hz at the tap classification band. The band was designed to be narrow and sharp, as the 80Hz peak was the most significant single spectral characteristic. The −3dB bandwidth of the tap classification band ranged from 0Hz to 172Hz and at 286Hz, the attenuation was −15dB.

The swipe resembled noise at a frequency band that ranged approximately from 70Hz to 1400Hz. The swipe classification band consisted of a band-pass filter that took into account the movement noise that ranged up to 300Hz. The cut-off frequencies were 384Hz and 1400Hz. The −3dB bandwidth of the swipe classification band ranged from 256Hz to 2144Hz.

Frequency responses of the two filter block can be seen in Figure 4.6.

4.2.2 Gesture duration measurement

Gesture durations were measured at the classification bands from the times that the noise gates remained open. A noise gate is an object that passes the signal only if the amplitude exceeds an opening threshold level. The closing threshold can be different from the opening threshold. This property is referred to as *hysteresis* [192] (see Figure 4.7). As the average RMS level for 60ms exceeded the open threshold level, the noise gate opened and triggered a timer. When the signal level attenuated below the closing threshold, the timer was triggered again and the gesture duration was recorded.

4.2.3 Classification logic

A tap gesture was determined to happen under the following conditions. If the 60ms RMS average value exceeds a certain threshold level at the tap gesture classification band, and the duration of the gesture is less than 60ms, the gesture is classified as a tap. The tap gesture threshold and the maximum duration time were iteratively calibrated.

The classification conditions for the swipe gesture followed a similar ruleset. If the 150ms RMS average at the swipe classification band exceeded a threshold level and the duration of the gesture was longer than 90ms, a swipe was considered to happen.

Further logic was included to consider the double tap and swipe, and the occasional contact taps. For a short period of time, the system was waiting for a second gesture of the same type. If two different gestures were detected inside this time period, the latter was selected. Also, if a swipe gesture was detected just after a tap (contact tap), the swipe was selected.

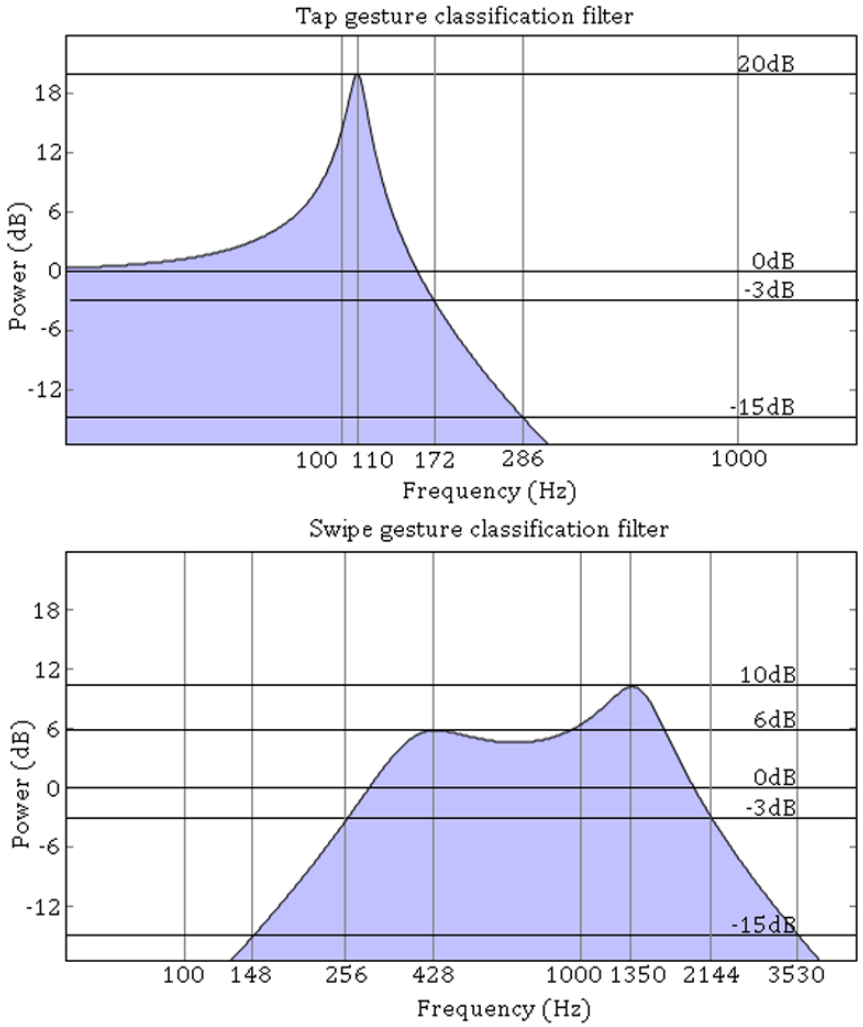

Figure 4.6: Frequency response curves of the classification band filters. Figure is exported from Max/MSP 5.

4.3 Auditory menu

A hierarchical auditory menu was constucted. The purpose of the menu was to evaluate the performance of the controller in the context of auditory interfaces. The selected sonification method was speech, so that the learning period would be very short. Besides speech, earcons were used to provide feedback sounds. As part of the menu design, the control to command mapping was created in a two phase user study.

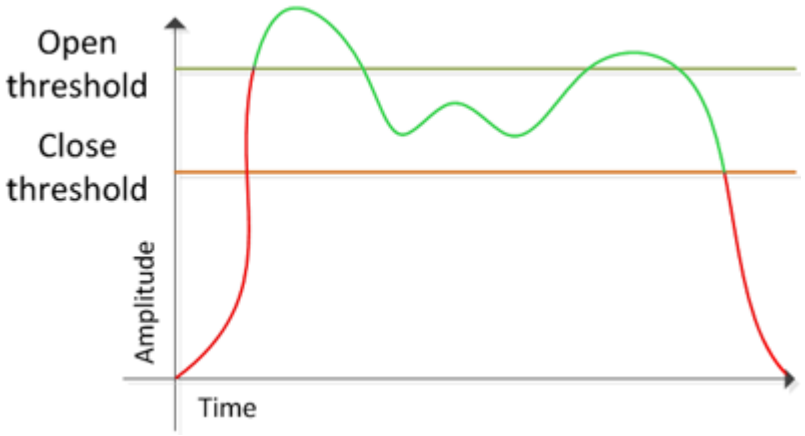

Figure 4.7: The behaviour of a noise gate with hysteresis. Green line depictes the signal that passes through the gate and red the portion of the signal that is cut off.

4.3.1 Description of the menu

The menu structure and contents were designed to be as simple and self-explatonary as possible. The object was that the user would be able to navigate the menu with only very the basic knowledge of the hierarchical structure. Therefore, very familiar categories and subcategories were selected as menu items.

The auditory menu contained three hierarchy levels. The first two levels contained two items and the third level six items. In total, the menu contained 2x2x6 items. The items were presented in alphabetic order and the first and last item in a menu level were linked. The first menu level includes plants and animals and the corresponding items on the second level there are cats, farm animals, flowers and fruits. The third level contains typical examples of each of the categories. The menu structure is presented in Appendix B.

4.3.2 Sound design

There were two types of sounds: item sounds and feedback sounds. All the sounds were monophonic. Item sounds were spoken English and they were generated using speech synthesis. The selected synthesis tool was FreeTTS 1.2 [193], which is based on the Flite speech synthesis engine. The default male voice settings were used. The user would hear once the sound of the item that he is currently pointing at i.e. the item that is under selection.

Earcons were used as feedback sounds. In general, feedback increases

the usability [100]. The earcons were generated inside Max/MSP using the standard MIDI libraries. When the user is moving up in hierarchy, two consecutive piano notes with rising pitch are played. Similarly, when the user moves down in hierachy, he will hear two consecutive notes with downward pitch. Furthermore, three high pitch notes are played when an item at the highest menu level is selected. Correspondingly, when the user is trying to go down in hierarchy at the root level, a dissonant chord is played. Dissonance is typically associated with something that is wrong or erroneous. The next item command produces a *click* sound and the previous item command a *swoop* sound. Feedback sounds and the corresponding conditions are presented in Figure 4.8.

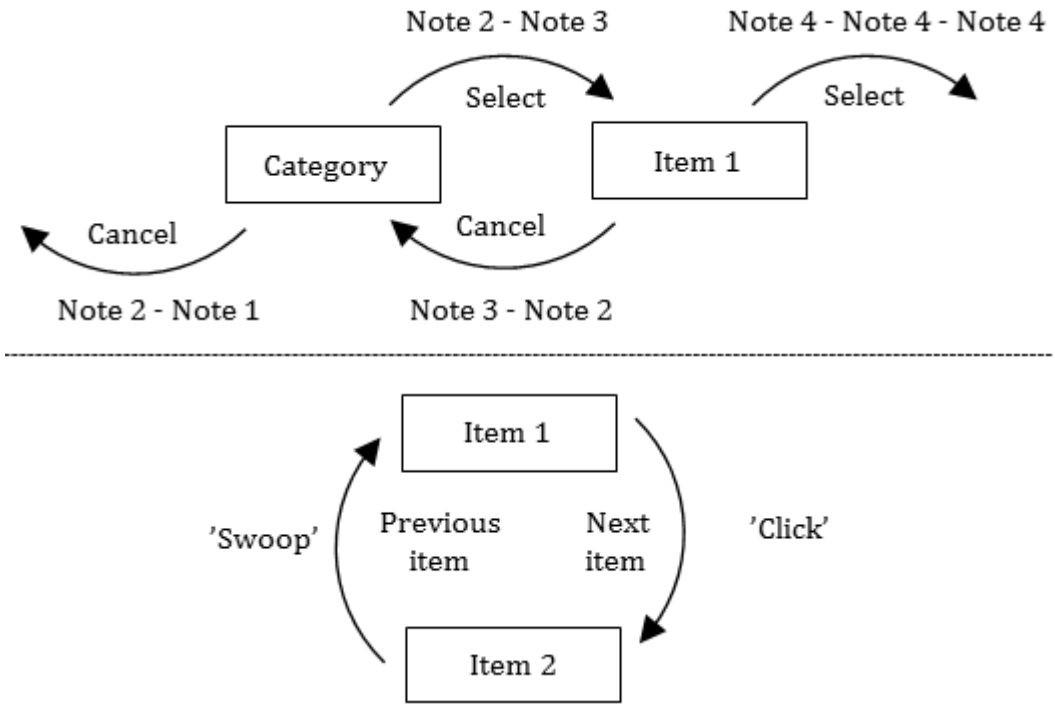

Figure 4.8: The feedback sounds and the corresponding auditory menu actions.

4.3.3 Gesture to command mapping

A two phase user study was conducted in order to find the most logical and natural semantic connections between the gestures and auditory menu commands. In total there were 24 possible mappings as there were four

gestures (tap, double tap, swipe, double swipe) and four commands (next item, previous item, ok, cancel). If the semantic connections are arbitrary and not thoroughly considered, the overall user experience becomes hindered [100].

Methodology

The first phase was a paper prototype. The controller and the auditory menu were excluded. Six students participated in the first phase.

First, the subjects were introduced the concept of navigating in an auditory menu using the four selected gestures. Then, the participants were asked to perform the gestures three times while sitting, standing and while walking. After this, the subjects had to connect each gesture to a logical menu command. As suggested in [100] for gesture to command mapping studies, the subjects were instructed that the task was to find the most natural connection for oneself. Furthermore, it was investigated whether the subjects had any problems in performing the selected gestures under any of the three conditions.

The second phase of the study included the prototype controller and the auditory menu. This phase was conducted one week after the first phase. The object of the second phase was to find the most generally preferable mapping out of the mappings found in the first phase. The second phase consisted of a training phase and a ranking phase.

Seven students participated the second phase. The study was conducted in Max/MSP programming environment. Sennheiser DT990 -headphones were used for audio playback and a 15" monitor was used to display a GUI. The GUI presented the current mapping and the subjects could change the mapping with a mouse. Furthermore, the gesture mappings were also printed into small pieces of paper.

During the training phase, the subjects were familiarized with the concept of auditory menu. The subjects were allowed to navigate the menu with a keyboard. Then the subjects were asked to insert the gesture controller into the pocket, in which they normally keep their PMD or mobile phone. The subjects had to successfully perform each gesture until the performance was fluent. The subjects were standing while performing the gestures.

In the ranking phase, the subjects tried and ranked each of the mappings. The order of the mappings was randomized. The subjects were asked to select three or more items in the auditory menu with each mapping. After this, the subjects ranked the mappings from best to worst according to their impressions. While ranking, the subjects were allowed to try the mappings again. Keywords given to the subjects for the ranking task were *natural*, *easy* and *intuitive*.

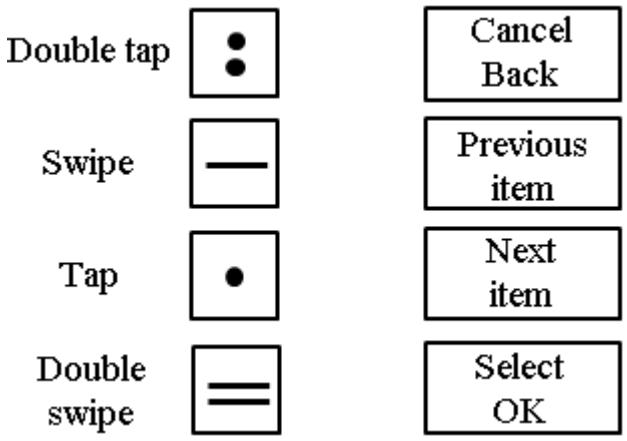

Figure 4.9: A diagram that was used in the gesture mapping study. The subjects had to connect each block to another.

Results

The subjects proposed three mappings in the first phase. One additional mapping was included by the author. The mappings are shown in Figure 4.10. There was a consensus on how the subjects created the semantic connections between gestures and commands. The single and the double version a gesture were connected to commands with the same category. For example swipe was suggested for *next item* and double swipe for *previous item*.

The first phase also revealed that the gestures are easy to perform. Three out of six subjects reported that they are familiar with or have devices that use touch screens.

Results for the second phase are presented in Table 4.3. The results show that the mapping **B** was the most preferable mapping for four subjects and the second most preferable for three subjects. The mappings **A** and **D** were quite even in popularity. The main difference between the mappings **B** and **D** is that **D** uses double tap for *Select/OK*, whereas **B** a single tap. Mapping **C** was the least preferable, most likely because it is using double gestures for *next item* and *Select/OK*, which are the most frequently used commands.

4.4 Traffic simulation experiment

The last development phase was to evaluate the performance of the controller. An experiment was conducted to simulate PMD usage scenario in traffic. The motivation of the eyes-free interaction is to increase the safety of the user of

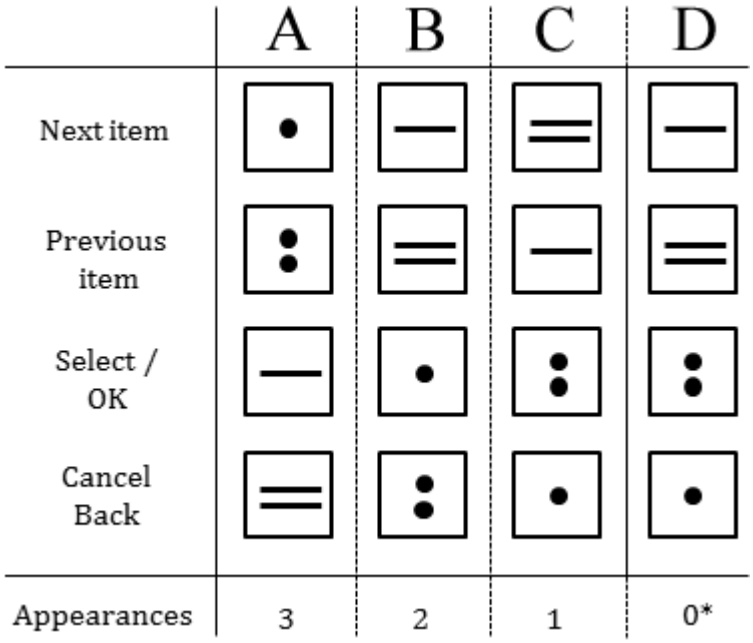

Figure 4.10: Proposed mappings from the first phase. The mapping D is proposed by the author.

	Subject						
Rank	1	2	3	4	5	6	7
1.	D	A	A	B	B	B	B
2.	B	B	B	D	D	D	A
3.	A	D	D	C	C	A	D
4.	C	C	C	A	A	C	C

Table 4.3: Gesture mappings derived from the second phase. Mapping B was found to be the most popular.

the PMD in traffic situations. The hypothesis is that the proposed controller coupled with an auditory interface provides enhanced capabilities to observe and react to sudden changes in the environment. Furthermore, the usability should not drastically diminish when moving from visual domain to auditory domain.

The traffic simulator experiment consisted of reaction time and task completion time measurements. The experiment followed a new methodology that was designed to evaluate the difference of an auditory interface and visual interface in terms of reaction times. Furthermore, the experiment evaluates

how much an environment tracking task affects task completion times on the two types of interfaces.

4.4.1 Methodology

Ten subjects, one female and nine male, aged 23-27 participated the experiment. The experiment was held in an acoustically treated listening room. One of the subjects reported tone deafness, but did not have problems in performing the experiment. Four subjects had some experience with the prototype controller and an auditory menu, as they had participated in the gesture mapping study.

Three interfaces, i.e. two auditory and a visual interface, were compared. The difference between the two auditory interfaces was that one was using the prototype controller and the other one a TV remote controller. Both were controlling the auditory menu presented in Section 4.3. A smartphone was used as a visual interface. The order of the interfaces was randomized for each subject.

The experiment consisted of an introduction, reaction time baseline measurement and the tests for each interface under two conditions. During the introduction, the subjects were motivated by explaining a scenario of using a PMD on a busy street and avoiding cars, bicycles and other obstacles. Furthermore, the subjects were familiarized with the concepts of auditory interfaces and the hierarchical menu structure. After the introduction the reaction time baseline was measured in an environment observation task.

Before each interface, a short training period was arranged. The subjects were familiarized with the interface and it was ensured that a subject was able to fully operate the menu. During the prototype controller training period, the subjects were required to perform a set of five flawless repetitions for each gesture. After a new interface was learned, the subjects had to practice using it by selecting three to five items in the menu.

The experiment lasted 30 - 45 minutes in total. The subjects were well focused during the whole experiment.

Experiment setup

The experiment was conducted in the Max/MSP 5 programming environment using a PC desktop computer, Philips SHC8585/100 wireless headphones, Impact USB Dance Pad, Hauppage WinTV Remote Control, HTC Desire HD mobile phone and two 15” LCD monitors.

The subjects were standing on the dance pad that was placed in the middle of the two monitors. The monitors were placed so that the subjects

had a sight line only to one monitor at a time without turning his or her head. A photo of the listening test setup can be seen in Figure 4.11.

Figure 4.11: The experiment setup for traffic simulator experiment. The dance pad was used to measure the reaction times to the stimulii that were displayed at either of the two screens at the sides.

Conditions

The first condition was a menu navigation task. The subject had to navigate the menu and select 15 items in a given randomized order. After the correct item was successfully selected, the subject was informed about the the next target item. In the case of the auditory interface, the next target was presented both aurally and visually on the monitors. In the case of the visual interface, the user navigated a corresponding folder structure in a smartphone touch screen using a file browser application. The target items were text files, that contained the next target. The selection intervals were measured.

The second condition added the environment observation task on top of the navigation task. Again, the subject had to navigate the menu in a given order, but in addition the subject had to continuously observe the two monitors. An arrow symbol was displayed at one of the monitors at

a random appearing interval of 2 - 10 seconds. The arrow could point to either up, down, left or right. When the subject detected an arrow, he had to react and press the corresponding arrow in the dance pad with his feet. The interval between the appearance of an arrow and the correct reaction was measured. The subjects were instructed that the observation task was the highest priority task. The subject was told that the arrow represents a car or another obstacle that could hit the subject.

4.4.2 Results

Reaction times

The reaction time results are shown in Figure 4.12. The bar represents the mean reaction time and the dotted lines the 95% confidence intervals for reaction times. Outliers were detected by using Grubbs's test with $\alpha = 0.05$. The mean baseline for reaction time was 1.70s. The mean reaction times for auditory interfaces were 1.88s for the remote controller and 2.08s for the gesture based prototype controller. The smartphone produced a 2.37s reaction time.

Selection intervals

The results for selection intervals are shown in Figure 4.13. The bar represents the mean selection intervals and the dotted lines the 95% confidence intervals for reaction times under the two conditions. Selection interval is the time between the selection of two correct items. The outliers were again detected with the Grubbs's test. The mean selection interval for the gesture contoller was 14.44s (condition 1) and 16.47s (condition 2) and for the remote controller 11.97s (condition 1) and 13.72s (condition 2). The selection intervals for phone were 7.02s (condition 1) and 9.51s (condition 2).

4.4.3 Discussion

The auditory interface produced a shorter reaction time than the visual interface. This is expected, as the subject did not have the need to divide the visual attention between the device screen and the environment. The remote controller was used as a reference control method for auditory interface. It provided slightly faster reaction times, which can yield that the prototype controller caused an increase in the cognitive load, which then reflected in the reaction time. As the gesture controller was a new method, the difference may diminish with further practice. However, the user needs to grasp the remote contoller whereas the gesture contoller can be kept inside a pocket.

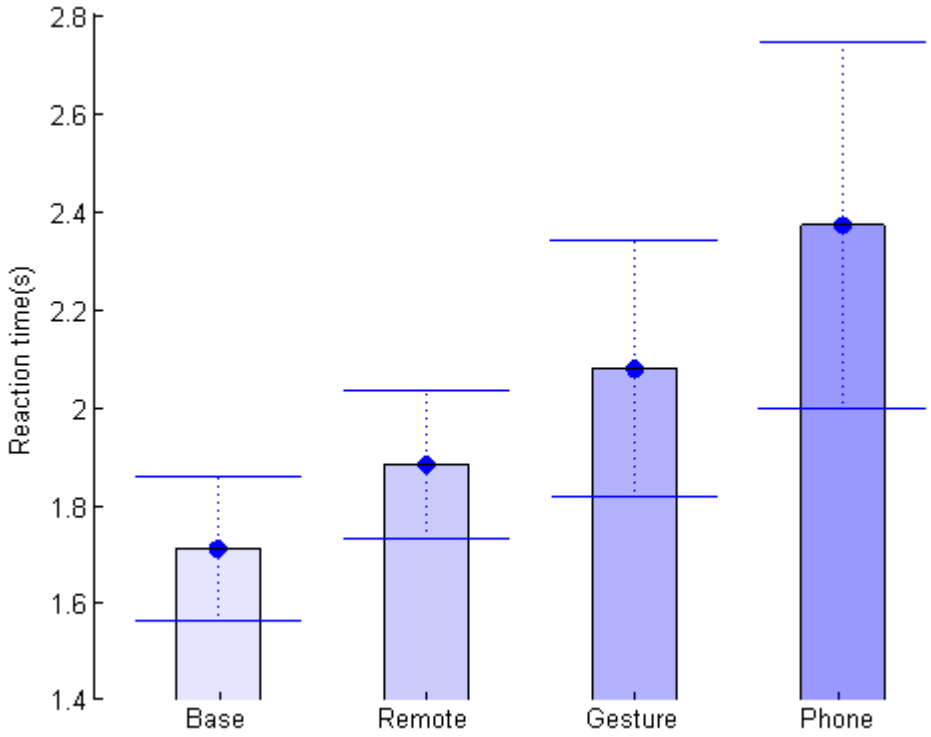

Figure 4.12: Reaction time results. The bars represent the reaction times and the dotted lines 95% confidence intervals.

The visual interface was a more rapid method for menu browsing. The subject could see all the items at the same time and directly point at the wanted item rather than passing by each item, which was the case in the auditory interfaces. This made the interface significantly more rapid but also very different by nature. However, the access times were not taken into account. For example, the access time for a mobile phone inside a pocket is approximately 4.6s [183]. Furthermore, for each interface, the selection intervals were slowed down under the condition 2. This increase was largest for the visual interface.

The gesture recognition system could be further developed. For example, instead of using a predefined classification, machine learning could be used to teach the classificator individual gestures.

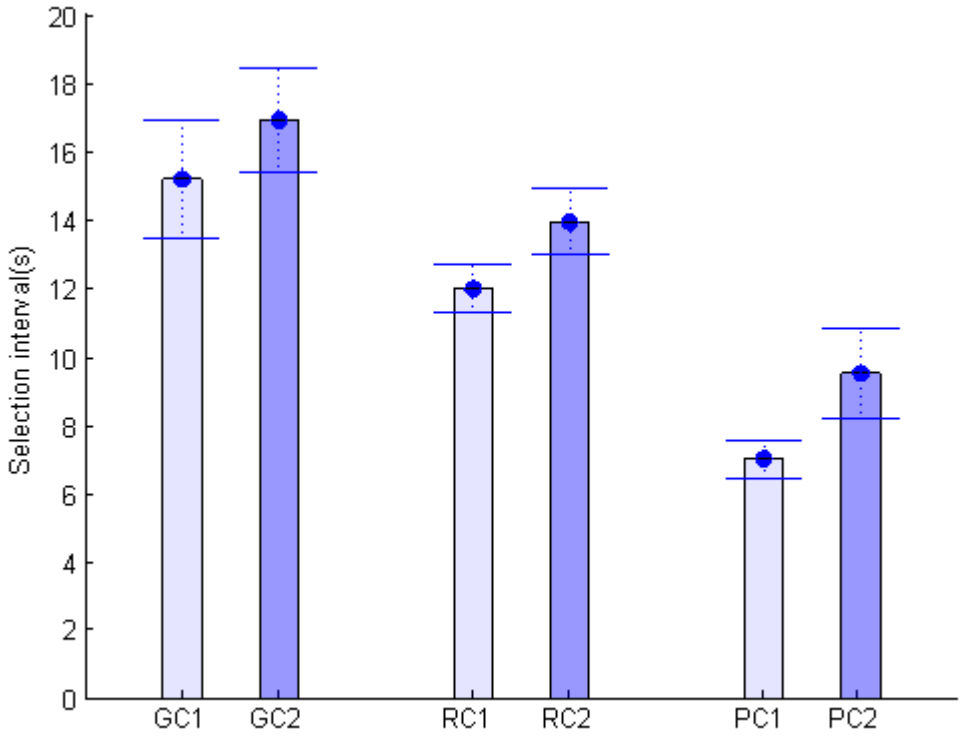

Figure 4.13: Mean selection intervals. GC1 = gesture ctrl. condition 1, GC2 = gesture ctrl. condition 2, RC1 = remote ctrl. condition 1, RC2 = remote ctrl. condition 2, PC1 = phone ctrl. condition 1, PC2 = phone ctrl. condition 2. The bars represent mean values and the dotted lines the 95% condifende intervals.

Chapter 5

Attention managing in auditory displays

This chapter addresses to the multitasking aspect of auditory interfaces. An approach is proposed that incorporates the conceptual ideas from ambient auditory displays and creates a virtual *auditory background* in a personal multilayered soundscape. The approach deals with auditory attention managing and attempts to solve how to present information from various sources so that...

- ...there is no cognitive overload?
- ...the auditory streams are distinguishable?
- ...the streams do not distract each other?

The definition of human multitasking is problematic. From a neurological perspective, it does not exist. The human is capable of performing only one cognitive task at a time. On the other hand the brain can switch attention from one task to another. This is the type of action that could be referred to as the human multitasking. On the downside, the brain requires time in selecting which task to currently perform and to change the task [194]. Luckily, the brain is cabable of learning to switch between tasks more efficiently [195, 196]. The proposed concept of multilayered personal soundscape utilizes this property of *divided attention* as a method for auditory multitasking. The aim is to produce a soundscape based auditory environment for efficient attention managing.

The structure of this chapter is as follows. First, the concept of a multilayer auditory interface and two hypothetical usage scenarios are presented. Then a two layer implementation is realized. Finally the concept is evaluated by measuring the speech intelligibility and subject impressions.

5.1 Multilayer auditory interface

To manage the auditory attention, the sounds generated by the various tasks are divided upon *sound layers*. The layers are prioritized and they are designed to be distinguishable from each other. Low priority layers are processed so that they are perceived as non-intrusive and non-distractive. High priority layers are designed to gain more user attention. It is intended that the auditory system recognizes the different layers as different auditory streams and that a priority level is induced into the sound. Furthermore, the method creates a setting for a virtual *cocktail party effect*.

The approach follows the metaphor of everyday listening. For example, as we listen to someone speak, we automatically neglect the environmental ambient sounds. This functions also to the other way around - we can focus attention to environmental ambient sounds and occasionally completely neglect the speech. If the sound stream has distinguishable characteristics, the auditory system is able to track a particular stream, even in a complex and noisy sound environment [197]. Such characteristics are for example the spatial location or timbre [197, 11].

Cocktail party effect has been investigated in auditory interfaces in [198, 199]. It was found that the spatial separation of sound sources improves the speech intelligibility of concurrent speakers. However, with three simultaneous speakers, the tracking becomes very hard and even distracting [200].

5.1.1 Auditory foreground and background

The simplest implementation of the concept is to divide the auditory space into a foreground and background. The tasks and the corresponding sound sources are classified into primary and secondary, or active and passive. Only one task is active at an instance and the other sources are classified as secondary, ambient sounds. The foreground is for the active task and the background layer is for monitoring passive tasks and processes. The user can then switch the attention from one layer to another. The users should also be able to change the priorities of the tasks i.e. move processes from one layer to another on the fly.

The background layer is processed so that the distraction is minimized and that the layer is distinguishable from the foreground. There are multiple possibilities in which the background layer can be processed, e.g. spectral properties, loudness and the use of spatial dimensions. As the aim is to create a background sound layer, the desirable perceptual properties are diffuseness, distance and softness - to create an impression of secondary, ambient sounds.

As the foreground layer is not processed, the auditory contrast between the layers is intended to highlight the active task.

The idea is further visualized Figure 5.1.

Figure 5.1: *"Evolution of auditory displays"*. Leftmost figure presents the monophonic presentation. Center figure presents the spatialization in which the sound objects are spatially separated in a form of a radial pie. This is the case in some modern spatial auditory interfaces e.g. in *EarPod* and *Hipui*. Rightmost figure presents the proposed method, where sound objects are separated and divided into auditory foreground and background.

5.1.2 Usage scenarios

In order to furthermore describe the multilayer auditory interface, two hypothetical usage scenarios are presented. It should to noted that the current work merely provides a concept for the interface designer and does not specify in details, how the interface should be designed.

Scenario 1 - Menu navigation and music

> The user is listening to music and intends to change the current song. He performs a rapid interaction with the device. The music track moves to the passive layer and an auditory menu appears to the active layer. The user navigates the auditory menu and each item is presented as text-to-speech sound object. Besides the auditory menu, the user hears the current track as if the band is playing on a distant location. As the user selects a new song, the music in the background changes and moves into the active layer.

Scenario 2 - Music and event feed

> The user is listening to a podcast. Meanwhile, a social networking application produces short notifications. These notifications are played as sound objects in the passive layer. One of the notification is particularly interesting to the user and he performs a short interaction. The podcast moves to the passive layer and pauses. The notification moves to the active layer and the complete message is presented to the user. After the message is finished, the podcast moves back to the active layer and continues.

5.1.3 Interaction

What is the ideal interaction method? More fundamentally, what kind of interaction the interface should be capable of providing? The interaction method should be designed on the basis of the interaction vocabulary, which is determined by the tasks and applications. Furthermore, the variety of different types of tasks is limited due to the auditory modality. Therefore, the interaction vocabulary and the corresponding interactions can remain simple.

The controller should provide methods for rapid interaction and basic operationality. The suggested minimum interaction level is that the user is able to enter *yes/no* type of commands and that the user is able to select a new active task from the set of secondary tasks. The gesture controller that was presented in the previous chapter could be one solution. It is particulary suitable, as it can be used through the clothes. Wearable proximity sensors are another interesting option. The hand proximity has an analogue to the spatial dimension of the interface. The hand proximity could for example control the auditory distance of an object and interchange the relative positions of active and passive tasks.

5.2 Two layer implementation

An implementation consisting of foreground and a one level background layer was realized. It was designed to present simultaneously speech and music. The background layer was produced by using binaural impulse responses (BRIRs). The intended spatial features of the BRIRs were the impression of distance and ambience. A block diagram of the implementation is shown in Figure 5.2. In the figure, the speech sources are in the background and music is in the foreground. As a simple interaction mechanism, the user was able to select between foreground and background layers. When the user pressed

spacebar, the foreground and background layers were switched by crossfading.

The current implementation has two sound sources and two layers. The number of background sources could be increased by measuring more BRIRs at different azimuth angles. Furthermore, the background depth could be varied by using BRIR filters that are measured from various distances. However, it is difficult to justify why to have more than one background depth level, except for layer transition effects. These effects may be used to improve the comprehension of the sound scape as a whole.

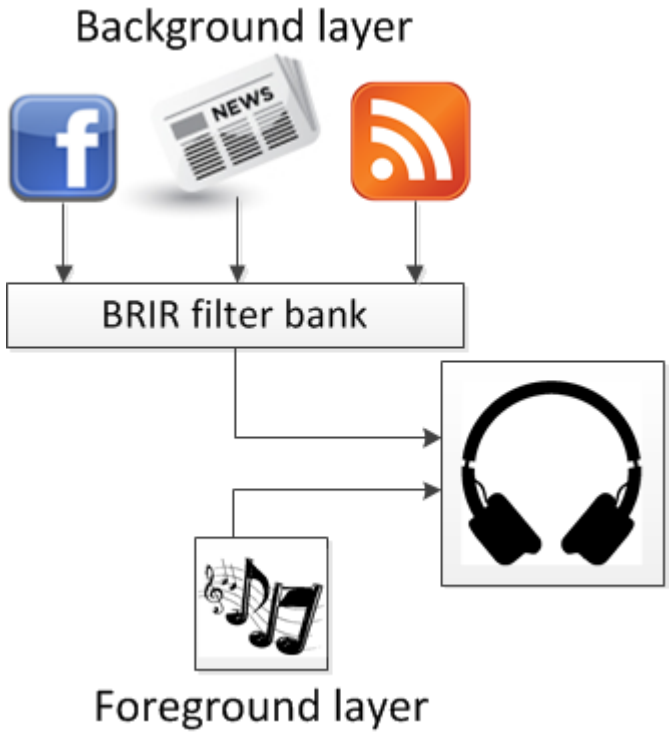

Figure 5.2: A simplified block diagram of the two layer implemenation. Here the music is in the foreground and speech sources are in the background.

5.2.1 Binaural impulse response measurements

Binaural impulse responses were measured both for speech and music. Two different locations and several measurement positions were used in order to find the most suitable BRIRs. The first location was a narrow laboratory room and the second a large storage room. Either rooms did not have any acoustic conditioning and they both were reverberant. The surface materials

were hard especially in the storage room, in which the floor and walls were concrete.

Equipment

The BRIRs were measured using a Neumann TU 81i dummyhead. The TU 81i uses KK 83 omni-directional condenser microphone capsules. The microphones were powered by a Neumann N 452 i power supply that was provided by the manufacturer. Furthermore, the excitation signal was played with a Genelec 8020B speaker. RME Hammerfall DSP multiface was used as the interface. The sampling frequency was 48kHz. The equipment is presented in Table 5.1. A block diagram of the measurement configuration is presented in Figure 5.3.

Item type	Manufacturer	Model
Dummy head	Neumann	TU 81 i
Power supply	Neumann	N 452 i
Speaker	Genelec	8020B
Desktop PC	Dell	Optiplex 745
Sound interface	RME	Hammerfall DSP multiface

Table 5.1: BRIR measurement equipment

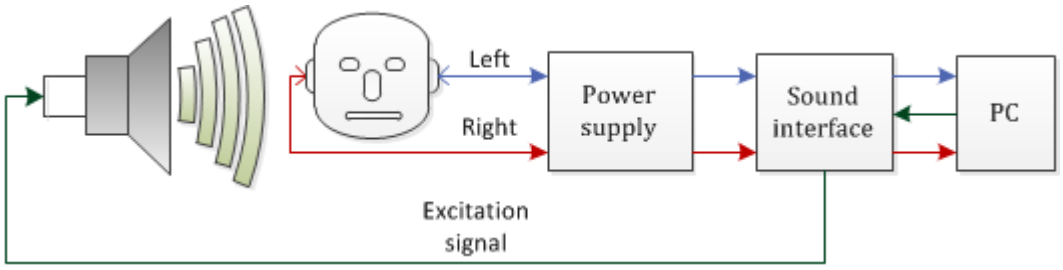

Figure 5.3: Block diagram of the BRIR measurement configuration

Measurements

The speaker and the dummyhead were horizontally and vertically aligned. Both were placed on a stand approximately one meter above the floor. The BRIRs were measured using a distance step of one meter. The distance ranged

from 1m to 7m in the laboratory room and from 7m to 11m in the storage room. Two azimuth angles were used. At $\theta = 0°$ the dummyhead was facing the loudspeaker and at $\theta = -45°$ the dummyhead was facing to the right of the loudspeaker. The measurement arrangements are shown in Figure 5.5.

The excitation signal was a logarithmic sweep. A spectrogram of the signal is presented at Figure 5.4.

The BRIRs

The binaural impulse responses were obtained by deconvoluting both of the recorded channels with the excitation signal. A BRIR for presenting music at the background was measured at $\theta = 0°$. The BRIR for speech was measured at $\theta = 45°$. The two selected BRIRs were measured in the storage room at the distance of 8m. This distance and location created the intended impression of ambient background sources. The two BRIRs are presented in Figure 5.6 and Figure 5.7. The BRIRs for speech was measured at $\theta = 45°$ in order to avoid the spatial overlapping of speech and lead vocals in music. The vocals are usually mixed to the center in stereo music [201].

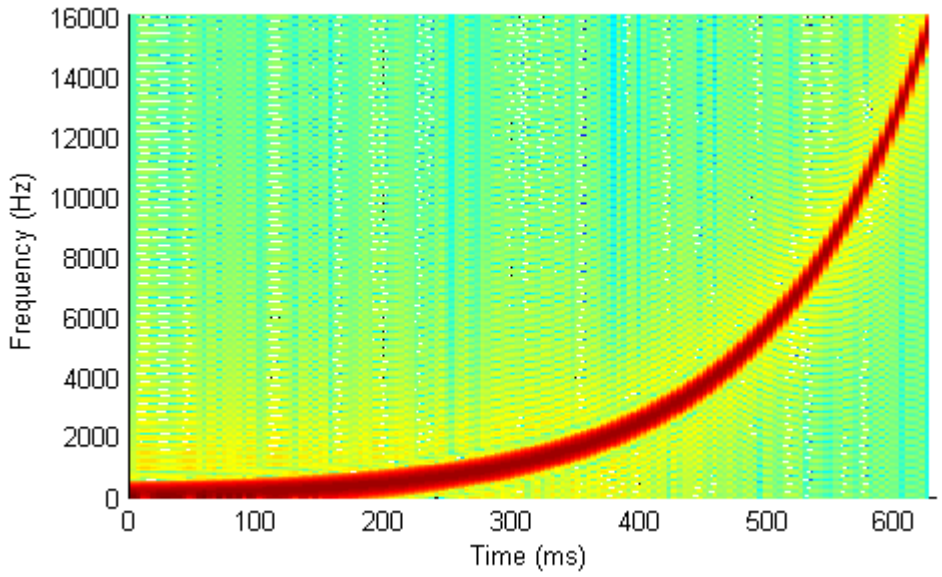

Figure 5.4: A logarithmic sine sweep was used as the excitation signal in the BRIR measurements.

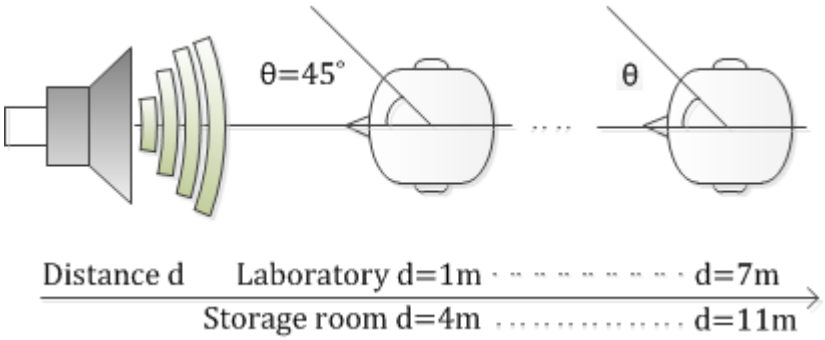

Figure 5.5: BRIR measurement arrangements. The measurements were performed in laboratory and a storage room. The sampling step was 1m at both locations. The BRIRs were measured at $\theta = 0°$ and at $\theta = 45°$.

5.2.2 Creating the layers

The sound scenes were produced using the recorded BRIRs and sound files consisting of speech and music. The sources on the background layer were convoluted with the corresponding BRIRs. Furthermore, the background layer was attenuated by 7dB with respect to the foreground layer to enhance the effect. Speech in the foreground layer was amplitude panned to the left to match the spatial position of speech in the background.

Energy balancing was performed in order to maintain the balance and control over the foreground and background layers. The average RMS energy and a scaling factor were calculated. Each track had equivalent RMS energy before applying 7dB attenuation to the background. The RMS energy for speech was calculated from the active speech parts. The RMS was calculated as

$$x_{RMS} = \sqrt{\frac{x_1^2 + x_2^2 + ... + x_n^2}{n}} \tag{5.1}$$

The scaling factor α was attained by

$$\alpha = \frac{x_{RMS,a}}{x_{RMS,b}} \tag{5.2}$$

5.3 Listening test

A listening test was conducted in which the BRIR filtering and amplitude panning methods were compared. Two quantitative measurements were performed. The first measure was a subjective speech intelligibility (SI)

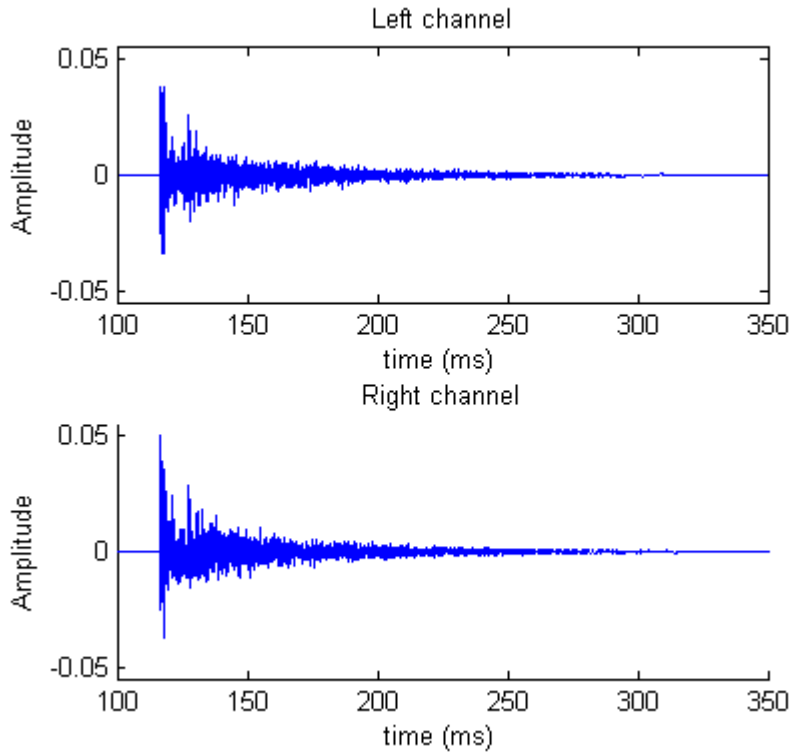

Figure 5.6: Waveform of the BRIR M, which was used for presenting background music. The BRIR Measured at the storage room at $d = 8$ meters, $\alpha = 0°$)

measurement. As the main goal of the layering concept is to efficiently provide information from various sources, SI is of great importance. The second measure was a the "Hedonic Utilitarian dimensions questionnaire" [202]. The questionnaire provides an overview of the subjects opinions and impressions, which are valuable at this point of early development.

5.3.1 Methodology

The listening test was conducted in a mixed Matlab and Max/MSP 5 environment. The system consisted of Beyerdynamic DT 990 headphones, a desktop computer, keyboard and a mouse. A total of 14 subjects participicated in the test.

The two layering methods, referred to as PAN and BRIR, were used to create two soundscapes (PAN1, PAN2, BRIR1, BRIR2). The two methods

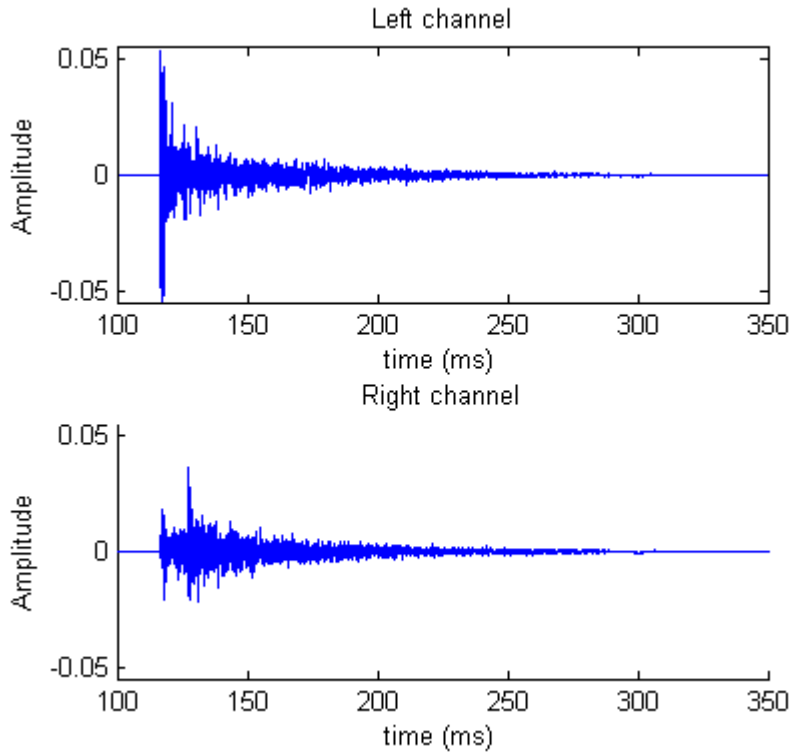

Figure 5.7: Waveform of the BRIR S, which was used for presenting backgroudn speech. The BRIR was measured at the storage room at $d = 8$ meters, $\alpha = 45°$)

differed in the way the background layer was created. In both methods, the background layer was attenuated by 7dB in comparison to the foreground. In the first method (soundscapes PAN1 and PAN2), the speech was amplitude panned to the front left. The second method (soundscapes BRIR1 and BRIR2) used the BRIR filtering. The soundscapes PAN1 and BRIR1 contained speech in the background and music in the foreground. The soundscapes PAN2 and BRIR2 contained speech in the foreground and music in the background. The sound source combinations used in the two soundscapes are presented in Table 5.2.

Speech intelligibility measurement

The speech intelligibility was measured using a modified Coordinate Response Measurement (CRM) [203] method. CRM has been used for measuring SI

	Soundscape			
	PAN1	**PAN2**	**BRIR1**	**BRIR2**
Foreground	Music	Speech	Music	Speech
Background	Speech	Music	Speech	Music

Table 5.2: The four soundscapes that were used in the listening test.

for example in multichannel and multitalker communications environments [18, 204, 205].

CRM contains a speech phrase corpus that has been constructed using eight different talkers. Each talker reads a phrase that contains three varying attributes. Each phrase is of the following format

Ready, *[codename]*, go to *[color] [number]*.

The possible contents for *[codename]*, *[color]* and *[number]* are presented in Table 5.3.

The CRM corpus was used to create speech sequences that consisted of 25 phrases. After each phrase there was a three second silent part. The structure of a speech sequence is presented in Table 5.4. The duration of a speech sequence was approximately two minutes. The corpus has been constructued by using multiple talkers, but only the talker number one was used in the current test.

Codenames	**Colors**	**Numbers**
Arrow	Blue	1
Baron	Green	2
Charlie	Red	3
Eagle	White	4
Hopper		5
Lager		6
Ringo		7
Tiger		8

Table 5.3: The possible attributes of a single phrase in the CRM corpus. Each phrase consisted of a codename, color and a number.

The subjects had to track a given codename. After hearing the codename, the subjects had to press a number that was associated with the phrase. A keyboard was used for input. In each speech sequence, there were five phrases at randomized positions that contained the correct codename.

Part					
1	2	3	4	...	50
Silence	Phrase #1	Silence	Phrase #2	...	Phrase #25

Table 5.4: Structure of a speech sequence. Each sequence contained 25 randomized phrases. Furthermore, five of the phrases contained the codename that was given to a subject.

During the tracking, the subjects were able to select their most preferred music by using the GUI. There were four music options, or *radio channels*, for each soundscape. This simulated a possible usage scenario where the user is listening to his favorite song. The user is thus primarily listening to music and only secondarily monitoring the background events or vice versa. The GUI is presented in Appendix D.

The HED/UT questionnaire

The impressions were evaluated with a Hedonic/Utilitarian (HED/UT) attributes questionnaire [202]. The questionnaire consists of 12 utilitarian and 12 hedonic statistically independent attributes. Each attribute is ranked on a 7 point Likert scale. The questionnaire was filled in after each soundscape. The subjects were asked to evaluate how each soundscape would apply to a real PMD use case.

Test structure

The test consisted of a training part and the four test cases (soundscapes PAN1, PAN2, BRIR1, BRIR2). During the training, the concept of the layered soundscapes in auditory interfaces was introduced to the subject. The training included a practice soundscape that was similar to an actual test case. The order of the test cases was randomized.

5.3.2 Results

Speech intelligibility

The speech intelligibility was evaluated from the number correct keypresses. The results are presented in Figure 5.8. The bars represent the mean value and the dotted lines the 95% confidence intervals. The mean SI was over 0.95 in PAN1, PAN2 and BRIR2. BRIR1 resulted in SI value of 0.7 and a larger variance.

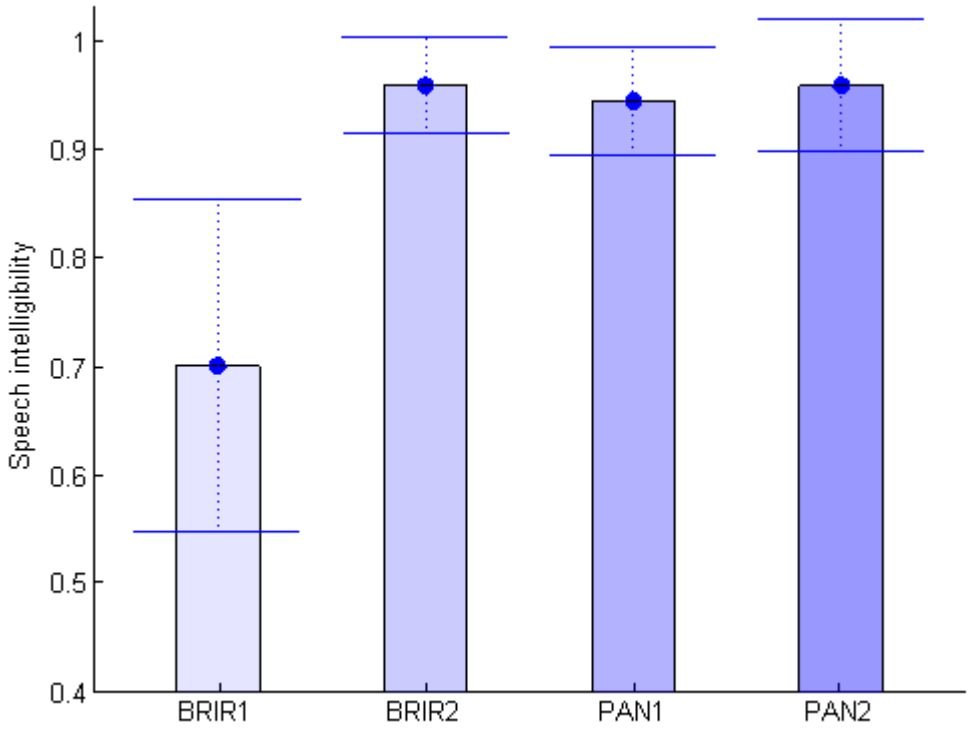

Figure 5.8: Speech intelligibility results. The bars denote the mean values and the dotted lines the 95% confidence intervals.

The HED/UT questionnaire

The questionnaire results are presented in Table 5.5. The questionnaire provided from neutral to positive results on both methods. Wilcoxon rank-sum test was performed for each pair (PAN1-BRIR1 and PAN2-BRIR2) on each attribute. The two methods differed statistically significantly ($p = 0.0386$) only on the attribute "thrilling" for the PAN2-BRIR2 pair.

5.3.3 Discussion

Speech intelligibility was decreased in the case in which speech was BRIR filtered. This result can be explained by the effect of the reverberation. Reverberation has a negative impact on the SI as it reduces the modulation depth of the speech [11]. Background sources did not decrease the SI under

the current conditions. This result suggests that although, D/R ratio is one of the primary cues in the sound source distance perception, the reverberation and should be produced more carefully.

The two methods were evaluated very similarly on the HED/UT scale. This may yield that the difference between the two methods was rather small, as there were only two simultaneous sound sources. Also, the current study did not represent a realistic usage scenario. Thus, future work should have an emphasis on the speech intelligibility and usage scenarios. For example personalized HRTFs and a more sophisticated distance model could be used. Usage scenarios should consider designing spatially profound hypothetical applications based on the auditory attention managing concept.

Utilitarian dimensions						
Attr.	PAN1	BRIR1	p	PAN2	BRIR2	p
EFE	3.9	4.3	.45	5.3	5.7	.34
HLP	3.7	4.1	.39	5.2	5.3	.60
FNC	3.8	4.5	.28	5.7	5.5	.88
NCS	3.8	4.1	.64	4.6	4.4	.71
PRC	4.1	4.0	.88	4.9	5.2	.42
BNF	4.0	4.3	.44	4.8	5.3	.26
USF	4.2	4.3	.98	5.1	5.4	.31
SNB	4.1	4.2	.71	4.9	5.4	.29
EFI	3.7	4.1	.45	5.3	5.2	.95
PRD	4.0	4.0	.90	4.9	4.6	.73
HND	3.7	4.9	.06	4.8	4.7	.98
Hedonic dimensions						
Attr.	PAN1	BRIR1	p	PAN2	BRIR2	p
FUN	4.5	4.7	.74	4.7	5.3	.18
EXC	4.5	4.5	1.0	4.5	5.1	.07
DGH	3.9	4.3	.46	4.3	5.0	.12
THR	3.8	4.3	.26	4.1	4.8	.04
ENJ	4.3	4.4	.78	4.4	5.1	.12
HPY	4.3	4.5	.86	4.9	5.2	.34
PLS	4.0	4.4	.43	5.0	5.3	.44
PLF	4.2	4.5	.41	4.7	5.1	.23
CHR	4.1	3.9	.65	4.5	4.7	.97
AMU	4.1	4.1	.95	4.8	4.9	.55
SNS	4.1	3.7	.27	4.4	4.5	.63
FNY	4.2	4.3	.89	4.4	4.7	.50

Table 5.5: The HED/UT questionnaire mean ratings and Wilcoxon Rank-sum test p-values. EFE = effective, HLP = helpful, FNC = functional, NCS = necessary, PRC = practical, BNF = beneficial, USF = useful, SNB = sensible, EFI = efficient, PRD = productive, HND = handy, FUN = fun, EXC = exciting, DGH = delightful, THR = thrilling, ENJ = enjoyable, HPY = happy, PLS = pleasant, PLF = playful, CHR = cheerful, AMU = amusing, SNS = senseous, FNY = funny.

Chapter 6

Rapid HRTF personalizing method

The last theme of the thesis considers HRTF personalization. This chapter presents a new method that is based on the concept of an aural pointer. The method is rapid and it can be implemented as an auditory game. Furthermore, besides the use of headphones, it does not require any additional equipment, which makes it ideal for personalization in PMDs.

The inclusion of spatial sound in auditory interfaces has many benefits. Spatial sound represents the most natural way of listening. It allows the creation new kinds of exciting interfaces. It also provides a method of mimicking the everyday listening, and furthermore methods for managing the auditory attention, as presented in Chapter 5. HRTFs provide a convenient method of producing spatial sound on PMDs. However, personalization is required in order to attain effective spatialization.

This chapter is structured as follows. First the concept of an aural pointer along previous work is presented. Then an aural pointer system is implemented. Finally, a spatial sound personalization experiment is conducted with the aural pointer.

6.1 The aural pointer

In spatial interfaces, it is often problematic how the user indicates the direction or interacts with the objects. There are various input devices ranging from 3D mice, data gloves, eye-tracking, mechanical, visual and acoustic tracking [206] and custom made special devices. Each of these provice different affordances and interaction metaphors. A common spatial interaction metaphor is pointing [206].

The problem of interacting with spatial objects is especially pronounced in spatial auditory interfaces. The sound objects may be located anywhere in

the auditory space. In the commonly used ring-topology, the user rotates the ring in order to move the item of interest to the front. The user will have to browse through each ring item on the way to the target item. The aural pointer attempts to implement the direct pointing metaphor in the auditory domain, so that the user does not need to, for example, rotate the menu.

The aural pointer is the auditory equivalent of the pointer in GUIs. It is a controllable virtual sound object that is able to indicate its current egocentric spatial position. The pointer emits sound that provides the lateralization cues. The user is therefore able to interact and point to the spatial objects i.e. virtual sound objects. Depending on the controller implementation, the user can move the pointer on a 2D-plane or in a 3D-space. The concept is illustrated in Figure 6.1. In the illustration, two sound sources, the pointer and a target item, are present on the horizontal plane.

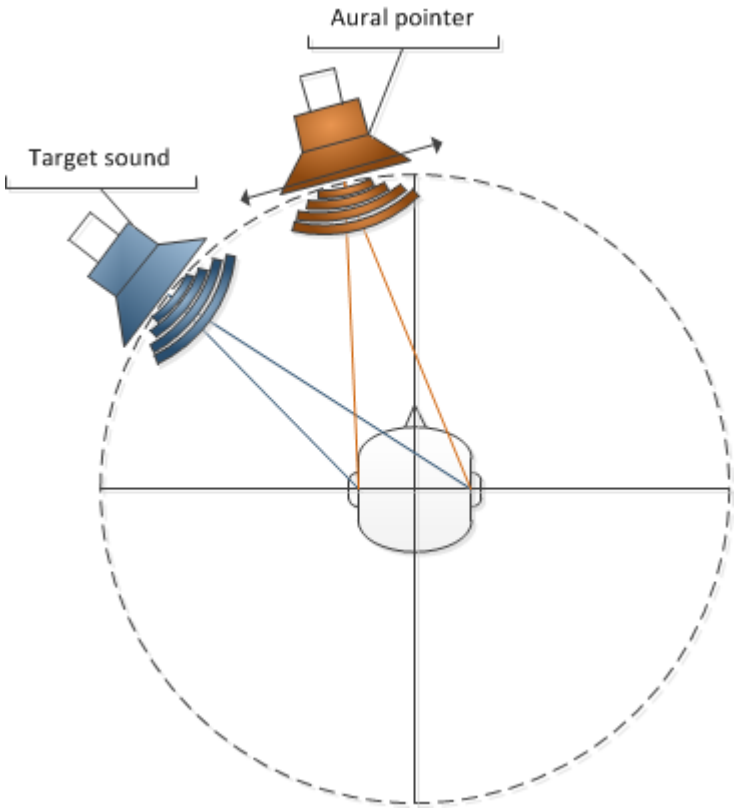

Figure 6.1: The concept of an aural pointer. The target item is a static sound source, but the spatial position of the pointer is controlled by the user. Both sources are spatially rendered.

6.2 Previous work on aural pointers

Auditory correspondents to visual pointers have been considered almost from introduction of the auditory interfaces. Nevertheless, the topic has gained very little attention and only few examples can be found.

An early aural pointer that resembled conceptually the GUI pointer was introduced in [207]. The work presented a speaker based spatial sound reproduction. A target sound was presented and the task was to move the aural pointer to point at the target. If the user clicked the right mouse button, he could hear the target sound. If the user concluded that the pointer is pointing at the target, he pressed the left mouse button to shoot it. Although the spatial sound was incorporated into the cursor as early as 1993 [207], it has been rarely used since. Another, more recent, example of the aural pointer concept was presented in [208]. Here the cursor is presented particularly as an spatial interaction method for virtual worlds. The cursor uses spherical coordinates (θ, ϕ) but also the pointer distance is contollable. The cursor was presented as a "hearcon", which is a spatial auditory object inside an auditory interaction realm (AIR) [209].

There have also been other kinds approaches. For example in a system called $\mathcal{SAGA}$ (Spatial Audio in Graphical Applications) [210] the pointer denoted the current listening point inside an auditory space. The user could hear the direction and distance to sources in the space and move the pointer accordingly. In [209, 211] a microphone metaphor is used. The closer the pointer was to a target item, the louder the target sound was.

6.3 An aural pointer system implementation

An aural pointer system for headphones was implemented in Matlab. The system was designed to present the aural pointer and a target item. The target item is an arbitrary auditory object with which the user is intended to interact and/or to point. The sources are located on the horizontal plane and the vertical dimension is neglected. The sound sources are spatialized by using HRTF filters from the CIPIC HRTF database [72].

6.3.1 Description of the system

The system consisted of the spatial renderer, keyboard and headphones. A block diagram of the system is presented in Figure 6.2.

The system renders two concurrent sound sources, the aural pointer and a target sound. The sounds are processed according to the two azimuth angles

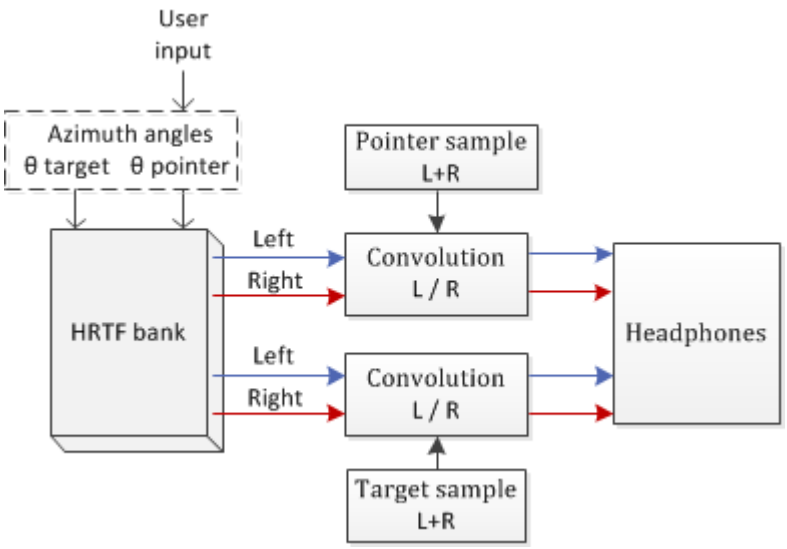

Figure 6.2: A block diagram of the aural pointer system.

θ_{pointer} and θ_{target}. The azimuth angle θ_{target} is defined by the system, but the θ_{pointer} is controllable by the user in real time. The possible azimuth values are determined by the HRTF sampling points.

In order to render a particular sound source, the system reads the azimuth angle set for the source, and selects the corresponding HRTF filter for left and right ear. The original monophonic sound source is convoluted with the two filters. Finally, the amplitude is normalized and the source is played through headphones.

The aural pointer is controlled with the arrow keys of a keyboard. Right arrow moves the pointer clockwise and left arrow counterclockwise. The pointer emits a short *beep* sound as it moves. The up arrow triggers a sound sample representing *shooting*. The shooting can be used for interaction e.g. selecting items. Furthermore, the ability to shoot gives a secondary lateralization cue to the user from which, he can assure that he is pointing at the intented direction. The pointer is audible only when the user performs an action - for example when the user moves the pointer or shoots. The target item is audible at all times.

6.3.2 The CIPIC HRTF database

The CIPIC HRTF Database (release 1.0) was used. The database contains HRTFs that have been measured from 45 subjects. The HRTFs have been

sampled at 25 azimuths and 50 elevation angles, which has resulted in a total of 1250 sampling points.

The azimuth values have not been uniformly sampled. The sampling steps are visualized in Figure 6.3. The angular increment between the steps has been 5° in the cyan, 10° in the green, 15° in the violet and 20° in the red sector. The highest sampling resolution was at the front ($\theta = \pm 45°$) and in the back ($\theta = 180° \pm 45°$). In total, there were 50 sampling points on the horizontal plane. The elevations were uniformly sampled by using an angular increment of 5.625°.

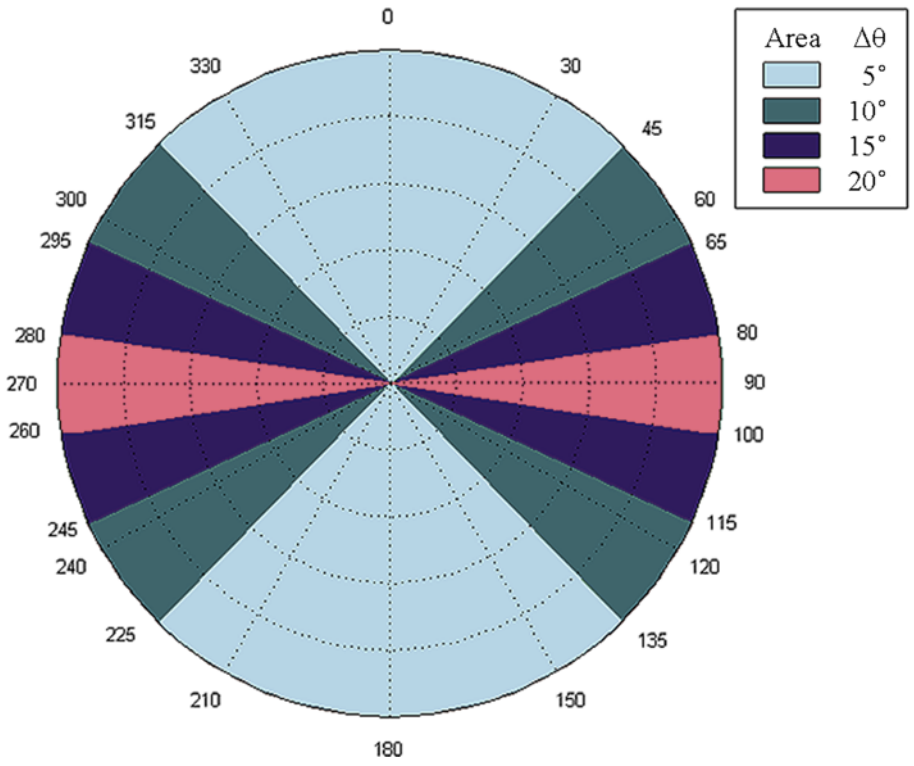

Figure 6.3: The CIPIC HRTF sampling resolution on the horizontal plane. The colored areas represent the used sampling step at different directions.

6.3.3 The sound samples

The pointer emitted two types of sounds: movement and selecting. The movement sound was a short beep. For selecting, there was a scifi-movie "laser" sound effect. The target objects are shot in the current implementation. The duration of the movement sample was 66ms. Furthemore, the sample

had a peak at 690Hz and above 1000Hz the spectrum had a decayed rapidly. The selecting sample had a duration of 260ms and a more evenly distributed frequency content that ranged up to 5200Hz. The two sound samples are presented in Figure 6.4 and Figure 6.5.

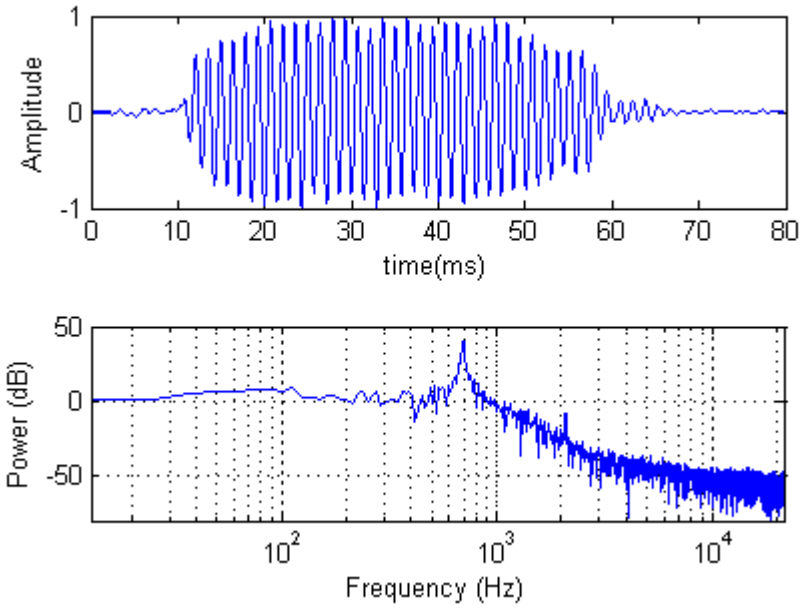

Figure 6.4: Waveform and spectrum of the cursor movement sample. The sample can be described as a "short beep".

6.4 HRTF personalization study with the aural pointer

A database matching personalization study was conducted. The objective was to find the most suitable HRTF set from the CIPIC HRTF Database for each subject by using an aural pointer system.

The study resembled an auditory version of a simple first person shooter game. The task of a subject was to aim and shoot at auditory targets. Then, the performance of a certain HRTF set was evaluated in terms of relative lateralization accuracy between the pointer and the target item. The hypothesis was that a greater accuaracy yields a better spatial resolution and thus better spatial reproduction. In the case of high accuracy on a certain

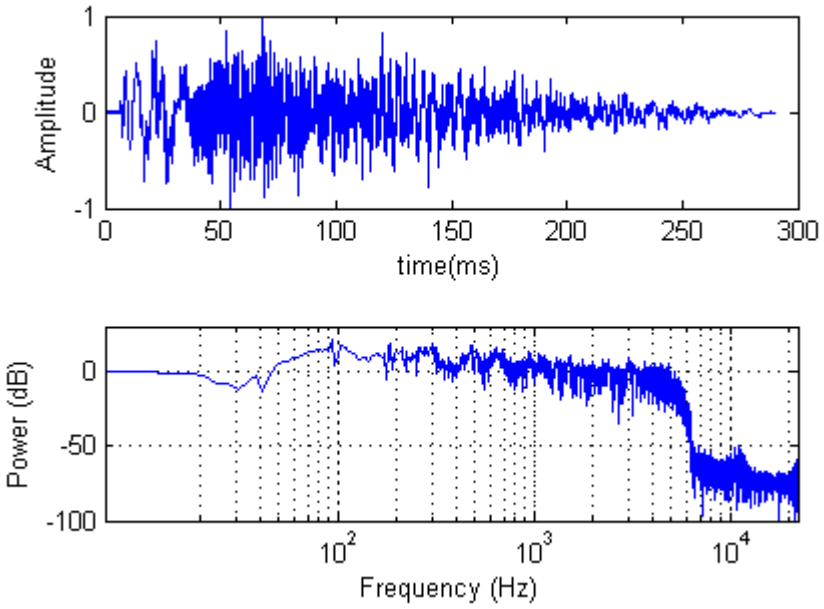

Figure 6.5: Waveform and spectrum of the sample that was used for shooting. The sample is a scifi type "laser" sound effect.

HRTF set, the subject is able to efficiently discriminate whether the two sounds are on the same half-plane (front or at the back) and how close the two sources are to each other. For example, the front-back confusions rapidly increase the lateralization error. Several HRTF sets were compared, and the one that produced the smallest lateralization error was selected as the candidate for a personalized HRTF set.

The proposed method is rapid, entertaining and does not require additional accessories. These properties makes it an ideal personalization method for PMDs.

6.4.1 Methodology

The listening test was conducted in Matlab. The system consisted of Beyerdynamic DT 990 headphones, a keyboard and a 15" monitor. The aural pointer implementation described in section 6.3 was used. The monitor displayed a GUI that provided information about the controls and the test progress. A total of 12 subjects participated in the test, one of which was considered to be an expert. The test lasted approximately 15 minutes.

Six HRTF sets from the standard CIPIC HRTF Database were selected into the test. Testing all of the 45 HRTFs would had been a very time consuming process, in which subject fatigue could not have been avoided. The selected HRTF sets were among a reduced CIPIC HRTF Database that had been constructed in [63]. The reduction was based on a measure of the highest spectral contrast between front and back locations at frequencies from 1000Hz to 10kHz [63]. This resulted in the selection of 13 HRTFs. In the current study, six out of the suggested 13 sets were selected randomly. These sets were numbered as subjects *008, 015, 021, 044, 119* and *154*.

Test structure

The test consisted of a training phase, a personalization phase and a validation phase. Each phase consisted of several cases. In each case the task was to use the keyboard to move the aural pointer to point at the same direction as the target sound and then shoot it. Shooting did not have any impact on the target - it was rather an indicator that provided a further lateralization cue to the subject. Furthermore, the subject could shoot as many times as he liked. As the subject was convinced that the two sound sources overlap spatially, he confirmed the direction by pressing the spacebar. The last pointing angle was saved and the test proceeded to the next case. The GUI is shown in Appendix E.

The training phase

During the training phase, the concept of an aural pointer was introduced to the subjects. The subjects practiced controlling the pointer with three practice targets. Data was not recorded in the training phase. As the subjects did not have anything to ask and they were familiar with the system, the actual test began.

The personalization phase

The purpose of the personalization phase was to find a candidate and a reference HRTF set. The candidate set is assumed to be the one that is closest to the individual HRTFs. The selection of the candidate sets was on the basis of the lateralization accuracy. The accuracy of the six HRTF sets were evaluated at three target angles. The target angles (θ_{target}) were 30°, 150° and 280°. The starting angle for the pointer (θ_{pointer}) was randomized for each target.

The target samples were speech. Speech is a familiar sound source that generally has a high lateralization accuracy. Three target samples were

randomly selected for each subject from a database of nine speech clips. The clips contained English and Dutch spoken by several people. Furthermore, the order of HRTF sets, target angles and samples were randomized. In total there were 18 test cases.

The average angular difference, i.e. the relative lateralization error, between the θ_{target} and θ_{pointer} was calculated for each HRTF set. The HRTF set that produced the smallest difference was selected as the candidate HRTF set. The set that produced the 2^{nd} largest error was selected as the reference HRTF set.

The validation phase

The candidate and the reference HRTF sets were compared in the validation phase. The purpose of the validation phase was to evaluate the difference of the two sets. The subject's task remained the same.

Eight target angles were randomly selected for both HRTF sets, which resulted in 16 test cases. The target sound was a sequence of beeps and sweeps. The sample contained a wide range of frequencies. Most of the energy was located at frequencies $200 - 2000$Hz.

The confusion classification

The front-back and back-front confusions were estimated from the results of the validation phase. A confusion was classified, if the pointer and the target were located at different quarters and if the angular difference between θ_{target} and θ_{pointer} exceeded a $35°$ threshold. This threshold was included in order to avoid the interpretation of a small lateralization error as front back confusion. The threshold corresponds to two pointer steps at the sides, where the human localization accuracy is poorest.

6.4.2 Results

Results for the personalization part are presented in Table 6.1. The table shows the candidate and reference HRTF set for each subject. The confusion ratios of the two HRTF sets from the validation phase are presented in Table 6.2. For eight subjects, the candidate HRTF set resulted in a diminished a confusion ratio. Two subjects encountered more confusion, and for two subjects there was no difference. Student's t-test was was performed on the estimated confusion ratios. The test produced a $p = 0.0163$. The average confusion ratio for the candidate sets was .385 and .573 for the reference sets.

[**?**]

HRTF	**Subject**											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Cand.	044	154	021	044	154	021	008	008	021	119	008	015
Ref.	021	119	008	154	021	154	021	015	119	154	044	021

Table 6.1: Personalization phase results. Each number assigned for candidate and reference sets correspond to the CIPIC HRTF Database subject number.

HRTF	**Subject**											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Cand.	0	.25	.50	.375	.125	.50	.25	.25	.875	.375	.625	.50
Ref.	.375	.50	.875	.75	.50	.375	.375	.625	.875	.375	.375	.875

Table 6.2: Front-back confusion ratios from the validation phase.

6.4.3 Discussion

The proposed personalization method was shown to reduce the front back confusion. Even though, the personalization phase was short, approximately 5 minutes, differences between the HRTF sets could be found. On the other hand only six sets were tested and there was a risk that a suitable for each subject set was not included.

The term *gamification* could be used to describe the method. Gamification refers to applying game thinking in a non-game context [212]. The current method turns the personalization into an entertaining process. However, the method did not incorporate common game elements such as feedback or rewarding mechanisms. The method could be implemented as a very polished and entertaining game. The more the users would play it, the more accurately the database match could made.

Future work should include more HRTF sets, target angles and an intelligent HRTF selection process. At times the subjects were spending an amount of time while they were carefully placing the aural cursor. Therefore, temporal element could be included into to force the subject to make rapid, intuitive, decisions. Another direction would be to move from HRTF database matching to, for example, HRTF modeling and to use the aural pointer, or some other subjective, perception based metric, to calibrate the system.

Chapter 7

Conclusions and future work

The initial objective of the thesis was increasing the safety of using a PMD in traffic. It is relevant for both pedestrians and drivers. The visual distraction and divided attention that occurrs during PMD usage increases the reaction times and inattentionall blindness, and reduces the situation awareness, which can lead to accidents. This thesis considered methods to replace the visual screen with a multitasking auditory interface.

Several fundamental issues are encountered when designing an auditory multitasking interface. The initial step was to investigate the consumer PMD usage habits by interviewing. Hypothetical consumer cases were developed based on the interviews. Then, the consumer cases were furthermore evaluated in a brainstorm session in order to find design concepts. Finally, the research topics were selected and implemented. The work was divided upon three distinct topics. The topics were eyes-free interaction, auditory multitasking and HRTF personalization in PMD context.

7.1 Eyes-free interaction

The interaction issues were tackled in Chapter 4. A gesture based eyes-free controller suitable for microinteractions was presented. The controller was based on acoustic classification of four tactile gestures. Furthermore, the controller is particularly attractive because it can be implemented virtually in any PMD that contains a microphone. The gestures can be performed for example through a pocket. The controller was used to control an auditory menu.

The prototype controller was compared against a visual interface in a reaction time experiment. It was found that the auditory modality reduces the reaction times. This is beneficial, as the reaction times have a correlation

to the traffic safety. Even more importantly, the auditory interface leaves the eyes free at all times. The user does not need to divide the visual attention between the device and the environment.

The gestural controller could be furthermore improved by including customization. The users could for example teach their individual gestures while standing, sitting and walking. Also, the classification system could be furthermore developed to detect for example the position of the tap and the direction of the swipe.

This study was presented at the Audio Engineering Society (AES) 134^{th} convention in Berlin [213].

7.2 Auditory multitasking

The auditory system is capable of analyzing the soundscapes with a high precision. We can essentially concentrate on one sound source, but we are also able to simultaneously monitor to the whole soundscape. A method was presented in Chapter 5 to present multiple simultaneous sound sources in a more natural soundscape setting. The approach takes advantage of the so called cocktail-party effect and the cognitive capability to switch attention between different tasks. Furthermore, it attempts to resemble more closely the real world sound events. Seldom, or never, do we listen only to one sound source at a time - ambient sounds are always present.

The multilayer auditory interface creates a soundscape that consists of several spatial depth layers. The idea is to provide efficient auditory environment for managing the attention and to improve the segregation of simultaneous sources. The method also attempts to increase the perception and comprehension of complex soundscapes by dividing it into foreground and background layers. The foreground is for the tasks with higher and the background layers for the tasks with lower priority.

The multilayer auditory interface is an attracting concept. Even though the current study did not find remarkable differences in the comparison of BRIR filtering and amplitude panning to create the layers, the concept should be furthermore developed in terms of more sophisticated sound design and interaction methods. It should be noted that currently the PMDs switch abruptly from one sound stream to another. Upon further development, it is crucial to consider the design guidelines for the possible applications and services that would support the concept. This leads to another paradigm of how to design new applications and how to adapt the existing applications into the auditory domain.

7.3 Rapid method for HRTF personalization

The use of spatial sound in auditory interfaces is beneficial. It enhances the sound source segregation and the spatial dimension enables new kinds of auditory interface design paradigms. As the functionalities and tasks have their own spatial location, the user can create efficient cognitive maps of the interface. Furthermore, the use of spatial sound can be an exciting new experience for the user.

A common headphone based spatialization method is the use of HRTFs. HRTFs describe the sound propagation from a point source to both ears. Each person has individual HRTFs and in general, the best spatialization is achieved by measuring the HRTFs individually. Non-individual HRTFs may cause front-back and back-front confusion, inaccurate lateralization and diminished spatial perception. Unfortunately, measuring the HRTFs of each PMD customer is practically an impossible task.

Chapter 6 presented a rapid HRTF personalization method. The method is suitable for PMD usage and it resembles an auditory game. It is based on the use of an aural pointer, which is a controllable auditory object. The idea behind the personalization method is that the perception controls the process. If the perception is distorted i.e. spatial resolution is poor, the system will try to find more suitable parameters. In practice, several CIPIC HRTF sets were tested for each person and the set that produced the smallest relative source position error between the aural pointer and a target item was determined to be the personalized HRTF set. The personalization method was found to reduce the front-back confusion ratios.

The method was presented at the Audio Engineering Society (AES) 132^{nd} convention in Budapest along with variations of the concept [214]. It was also patented under the number WO2013064943.

7.4 Final thoughts

Auditory multitasking in personal media devices is a complex mixture of multiple disciplines including acoustics, signal processing, psychoacoustics, cognitive sciences, interaction design and sound design. The current work considered some aspects of it but much is left for future work. Hopefully, the spatial sound could one day be as an integral element of the auditory interfaces and we would see some truly functional and fascinating new interfaces.

Bibliography

[1] Deloitte consumer review, beyond the hype: The true potential of mobile, 2013.

[2] David C. Schwebel, Despina Stavrinos, Katherine W. Byington, Tiffany Davis, Elizabeth E. O'Neal, and Desiree de Jong. Distraction and pedestrian safety: How talking on the phone, texting, and listening to music impact crossing the street. *Accident Analysis and Prevention*, 45:266–271, 2012.

[3] Jack Nasar, Peter Hecht, and Richard Wener. Mobile telephones, distracted attention, and pedestrian safety. *Accident Analysis and Prevention*, 40(1):69–75, 2008.

[4] S. L. Chisholm, Jeff K. Caird, and J. Lockhart. The effects of practice with MP3 players on driving performance. *Accident Analysis & Prevention*, 40(2):704–713, 2008.

[5] Simon G. Hosking, Kristie L. Young, and Michael A. Regan. The effects of text messaging on young drivers. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 51(4):582–592, 2009.

[6] Kristie Young, Michael Regan, and M. Hammer. Driver distraction: A review of the literature. *Distracted driving. Sydney, NSW: Australasian College of Road Safety*, pages 379–405, 2007.

[7] Antti Oulasvirta. The fragmentation of attention in mobile interaction, and what to do with it. *interactions*, 12(6):16–18, 2005.

[8] Antti Oulasvirta, Sakari Tamminen, Virpi Roto, and Jaana Kuorelahti. Interaction in 4-second bursts: the fragmented nature of attentional resources in mobile HCI. In *Proc. of the SIGCHI conference on Human factors in computing systems (CHI '05)*, pages 919–928, New York, NY, USA, 2005.

[9] Turkka Keinonen. User-centered design and fundamental need. In *Proc. of the* 5^{th} *Nordic conference on Human-computer interaction: building bridges*, pages 211–219, Lund, Sweden, 2008.

[10] Jens Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. M.I.T. Press, Cambridge, MA, 1997.

[11] Matti Karjalainen. *Kommunikaatioakustiikka*. Aalto University, School of Electrical Engineering, Department of Signal Processing and Acoustics, Espoo, Finland, 2009.

[12] Sharaf Hameed, Jyri Pakarinen, Kari Valde, and Ville Pulkki. Psychoacoustic cues in room size perception. In *Proc. of the* 116^{th} *Audio Engineering Society Convention (AES)*, Berlin, Germany, 2004.

[13] Claudia Carello, Krista L. Anderson, and Andrew J. Kunkler-Peck. Perception of object length by sound. *Psychological science*, 9(3):211–214, 1998.

[14] Juan G. Roederer. *The physics and psychophysics of music: an introduction*. Springer Publishing Company, Incorporated, 2008.

[15] Durand R. Begault. *3-D sound for virtual reality and multimedia*. National Aeronautics and Space Administration, Ames Research Center, 2000.

[16] Matti Gröhn. *Application of spatial sound reproduction in virtual environments: experiments in localization, navigation, and orientation*. Helsinki University of Technology, 2006.

[17] Aki Härmä, Julia Jakka, Miikka Tikander, Matti Karjalainen, Tapio Lokki, Jarmo Hiipakka, and Gaëtan Lorho. Augmented reality audio for mobile and wearable appliances. *Journal of the Audio Engineering Society*, 52(6):618–639, 2004.

[18] Ville Pulkki. Spatial sound reproduction with directional audio coding. *Journal of the Audio Engineering Society*, 55(6):503–516, 2007.

[19] B. Shinn-Cunningham. Applications of virtual auditory displays. In *Proc. of the* 20^{th} *International Conference of the IEEE Engineering in Biology and Medicine Society*, volume 3, pages 1105–1108, Hong Kong, China, 1998.

[20] Thomas D. Rossing, F. Richard Moore, and Paul A. Wheeler. *The science of sound*, volume 3. Addison-Wesley Massachusetts., 2002.

[21] Lloyd A. Jeffress. A place theory of sound localization. *Journal of comparative and physiological psychology*, 41(1):35–39, 1948.

[22] Richard M. Stern, Guy J. Brown, and DeLiang Wang. Binaural sound localization. *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, pages 147–185, 2006.

[23] Simon Carlile. *The physical and psychophysical basis of sound localization*, pages 27–78. Springer Berlin Heidelberg, 1996.

[24] Henrik Møller and Daniela Toledo. The role of spectral features in sound localization. In *Proc. of the* 124^{th} *Audio Engineering Society Convention (AES)*, Amsterdam, The Netherlands, 2008.

[25] William G. Gardner. Spatial audio reproduction: Towards individualized binaural sound. *The Bridge*, 34(4):37–42, 2004.

[26] V. Ralph Algazi, Richard O. Duda, Ramani Duraiswami, Nail A. Gumerov, and Zhihui Tang. Approximating the head-related transfer function using simple geometric models of the head and torso. *The Journal of the Acoustical Society of America*, 112:2053–2064, 2002.

[27] V. Ralph Algazi, Richard O. Duda, Reed P. Morrison, and Dennis M. Thompson. Structural composition and decomposition of HRTFs. In *Proc. of the IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 103–106, New Platz, NY, USA, 2001.

[28] Paul M. Hofman, Jos G. A. Van Riswick, and A. John Van Opstal. Relearning sound localization with new ears. *Nature neuroscience*, 1(5):417–421, 1998.

[29] CIPIC HRTF database files, release 1.0, august 15, 2001.

[30] Allen William Mills. On the minimum audible angle. *The Journal of the Acoustical Society of America*, 30(4):237–246, 1958.

[31] Frederic L. Wightman and Doris J. Kistler. The dominant role of low-frequency interaural time differences in sound localization. *The Journal of the Acoustical Society of America*, 91(3):1648–1661, 1992.

[32] G. Bruce Henning. Detectability of interaural delay in high-frequency complex waveforms. *The Journal of the Acoustical Society of America*, 55(1):84–90, 1974.

[33] Gösta Ekman. Weber's law and related functions. *The Journal of Psychology*, 47(2):343–352, 1959.

[34] Robert S. Woodworth. *Experimental psychology.* New York: Holt, Rinehart & Winston, 1938.

[35] Erno H. A. Langendijk and Adelbert W. Bronkhorst. Contribution of spectral cues to human sound localization. *The Journal of the Acoustical Society of America*, 112(4):1583–1596, 2002.

[36] Alan D. Musicant and Robert A. Butler. The influence of pinnae-based spectral cues on sound localization. *The Journal of the Acoustical Society of America*, 75(4):1195–1200, 1984.

[37] George F. Kuhn. *Physical acoustics and measurements pertaining to directional hearing*, pages 3–25. Springer, 1987.

[38] Frederic L. Wightman and Doris J. Kistler. Resolution of front-back ambiguity in spatial hearing by listener and source movement. *The Journal of the Acoustical Society of America*, 105(5):2841–2853, 1999.

[39] Willard R. Thurlow and Philip S. Runge. Effect of induced head movements on localization of direction of sounds. *The Journal of the Acoustical Society of America*, 42(2):480–488, 1967.

[40] Yukio Iwaya, Yôiti Suzuki, and Daisuke Kimura. Effects of head movement on front-back error in sound localization. *Acoustical science and technology*, 24(5):322–324, 2003.

[41] David R. Perrott and Kourosh Saberi. Minimum audible angle thresholds for sources varying in both elevation and azimuth. *The Journal of the Acoustical Society of America*, 87(4):1728–1731, 1990.

[42] Donald H. Mershon and L. Edward King. Intensity and reverberation as factors in the auditory perception of egocentric distance. *Perception & Psychophysics*, 18(6):409–415, 1975.

[43] Pavel Zahorik. Assessing auditory distance perception using virtual acoustics. *The Journal of the Acoustical Society of America*, 111:1832–1846, 2002.

[44] Paul D. Coleman. Failure to localize the source distance of an unfamiliar sound. *The Journal of the Acoustical Society of America*, 34(3):345–346, 1962.

[45] Pavel Zahorik, Douglas S. Brungart, and Adelbert W. Bronkhorst. Auditory distance perception in humans: A summary of past and present research. *Acta Acustica united with Acustica*, 91(3):409–420, 2005.

[46] Peter McGregor, Andrew G. Horn, and Melissa A. Todd. Are familiar sounds ranged more accurately? *Perceptual and motor skills*, 61(3f):1082, 1985.

[47] Georg von Békésy. The moon illusion and similar auditory phenomena. *The American journal of psychology*, 62(4):540–552, 1949.

[48] Yan-Chen Lu and Martin Cooke. Binaural distance perception based on direct-to-reverberant energy ratio. In *Proc. of the International Workshop Acoustic Echo and Noise Control (IWAENC)*, Washington, USA, 2008.

[49] Yan-Chen Lu and Martin Cooke. Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(7):1793–1805, 2010.

[50] Paul D. Coleman. Dual role of frequency spectrum in determination of auditory distance. *The Journal of the Acoustical Society of America*, 44(2):631–632, 1968.

[51] Alex D. Little, Donald H. Mershon, and Patrick H. Cox. Spectral content as a cue to perceived auditory distance. *Perception*, 21(3):405–416, 1992.

[52] Uno Ingård. A review of the influence of meteorological conditions on sound propagation. *The Journal of the Acoustical Society of America*, 25(3):405–411, 1953.

[53] N. I. Durlach, A. Rigopulos, X. D. Pang, W. S. Woods, A. Kulkarni, H. S. Colburn, and E. M. Wenzel. On the externalization of auditory images. *Presence: Teleoperators and Virtual Environments*, 1(2):251–257, 1992.

[54] Frederic L. Wightman and Doris J. Kistler. Factors affecting the relative salience of sound localization cues. *Binaural and spatial hearing in real and virtual environments*, 1:1–23, 1997.

[55] Hyun Jo, William Martens, and Youngjin Park. Evaluating candidate sets of head-related transfer functions for control of virtual source elevation. In *Proc. of the* 40^{th} *Audio Engineering Society Conference (AES)*, Tokyo, Japan, 2010.

[56] Elizabeth M. Wenzel, Frederic L. Wightman, and Doris J. Kistler. Localization with non-individualized virtual acoustic display cues. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI '91)*, pages 351–359, New Orleans, LA, USA, 1991.

[57] Andreas Silzle. Selection and tuning of HRTFs. In *Proc. of the* 112^{th} *Audio Engineering Society Convention (AES)*, Munich, Germany, 2002.

[58] Henrik Møller, Michael Friis Sørensen, Clemen Boje Jensen, and Dorte Hammershøi. Binaural technique: Do we need individual recordings? *Journal of the Audio Engineering Society*, 44(6):451–469, 1996.

[59] Song Xu, Zhizhong Li, and Gaviriel Salvendy. *Individualization of head-related transfer function for three-dimensional virtual auditory display: a review*, pages 397–407. Springer, 2007.

[60] Dmitry N. Zotkin, Ramani Duraiswami, and Larry S. Davis. Creation of virtual auditory spaces. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages II–2113–II–2116, Orlando, FL, USA, 2002.

[61] Dmitry N. Zotkin, Ramani Duraiswami, and Larry S. Davis. Rendering localized spatial audio in a virtual auditory space. *Multimedia, IEEE Transactions on*, 6(4):553–564, 2004.

[62] Durand R. Begault, Elizabeth M. Wenzel, and Mark R. Anderson. Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. *Journal of the Audio Engineering Society*, 49(10):904–916, 2001.

[63] Agnieszka Roginska, Gregory H. Wakefield, and Thomas S. Santoro. User selected HRTFs: Reduced complexity and improved perception. In *Proc. of the Undersea Human Systems Integration Symposium (UHSI)*, Providence, RI, USA, 2010.

[64] Francis Rumsey. Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm. *Journal of the Audio Engineering Society*, 50(9):651–666, 2002.

[65] Elizabeth M. Wenzel and Scott H. Foster. Perceptual consequences of interpolating head-related transfer functions during spatial synthesis. In *Proc. of the IEEE Workshop on Applications of Signal Processing to*

Audio and Acoustics (WASPAA), pages 102–105, New Paltz, NY, USA, 1993.

[66] Takashi Takeuchi, Philip A. Nelson, Ole Kirkeby, and Hareo Hamada. Influence of individual head-related transfer function on the performance of virtual acoustic imaging systems. In *Proc. of the* 104^{th} *Audio Engineering Society Convention (AES)*, Amsterdam, The Netherlands, 1998.

[67] Piotr Majdak, Peter Balazs, and Bernhard Laback. Multiple exponential sweep method for fast measurement of head-related transfer functions. *Journal of the Audio Engineering Society*, 55(7/8):623–637, 2007.

[68] Klaus A. J. Riederer. Repeatability analysis of head-related transfer function measurements. In *Proc. of the* 105^{th} *Audio Engineering Society Convention (AES)*, San Francisco, CA, USA, 1998.

[69] Olivier Warusfel. LISTEN HRTF database, http://recherche.ircam.fr/equipes/salles/listen/, 2003.

[70] Ville Pulkki, Mikko-Ville Laitinen, and Ville Sivonen. Hrtf measurements with a continuously moving loudspeaker and swept sines. In *Proc. of the* 128^{th} *Audio Engineering Society Convention (AES)*, 2010.

[71] Javier Gómez Bolaños and Ville Pulkki. Hrir database with measured actual source direction data. In *Proc. of the* 133^{th} *Audio Engineering Society Convention (AES)*, 2012.

[72] V. Ralph Algazi, Richard O. Duda, Dennis M. Thompson, and Carlos Avendano. The CIPIC HRTF database. In *Proc. of the IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 99–102, New Platz, NY, USA, 2001.

[73] Henrik Møller, Michael Friis Sørensen, Dorte Hammershøi, and Clemen Boje Jensen. Head-related transfer functions of human subjects. *Journal of the Audio Engineering Society*, 43(5):300–321, 1995.

[74] Klaus A. J. Riederer. Part IVa: Effect of cavum conchae blockage on human head-related transfer functions. In *Proc. of the* 18^{th} *Int. Congress on Acoustics (ICA)*, pages I–787 – I–790, Kyoto, Japan, 2004.

[75] Masayuki Morimoto and Hitoshi Aokata. Localization cues of sound sources in the upper hemisphere. *Journal of the Acoustical Society of Japan (E)*, 5(3):165–173, 1984.

[76] James C. Makous and John C. Middlebrooks. Two-dimensional sound localization by human listeners. *The Journal of the Acoustical Society of America*, 87(5):2188–2200, 1990.

[77] Pauli Minnaar, Jan Plogsties, and Flemming Christensen. Directional resolution of head-related transfer functions required in binaural synthesis. *Journal of the Audio Engineering Society*, 53(10):919–929, 2005.

[78] Lauri Savioja, Jyri Huopaniemi, Tapio Lokki, and Ritta Väänänen. Creating interactive virtual acoustic environments. *Journal of the Audio Engineering Society*, 47(9):675–705, 1999.

[79] Luiz W. P. Biscainho, Fabio P. Freeland, and Paulo Sergio Ramirez Diniz. Using inter-positional transfer functions in 3D-sound. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages II–1961–II–1964, Orlando, FL, USA, 2002.

[80] Gerald Enzner. 3d-continuous-azimuth acquisition of head-related impulse responses using multi-channel adaptive filtering. In *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA'09. IEEE Workshop on*, pages 325–328, 2009.

[81] Martin Rothbucher, Tim Habigt, Julian Habigt, Thomas Riedmaier, and Klaus Diepold. Measuring anthropometric data for HRTF personalization. In *Proc. of the* 6^{th} *International Conference on Signal-Image Technology and Internet-Based Systems (SITIS)*, pages 102–106, Kuala Lumpur, Malesia, 2010.

[82] Navarun Gupta, Armando Barreto, Manan Joshi, and Juan Carlos Agudelo. HRTF database at FIU DSP lab. In *Proc. of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 169–172, Dallas, TX, USA, 2010.

[83] Bill Gardner and Keith Martin. HRTF measurements of a KEMAR dummy-head microphone. *Massachusetts Institute of Technology*, 280:1–7, 1994.

[84] DSPeaker. *DSPeaker, HeaDSPeaker manual*. 2010.

[85] Dmitry N. Zotkin, Jane Hwang, R. Duraiswaini, and Larry S. Davis. HRTF personalization using anthropometric measurements. In *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 157–160, New Paltz, NY, USA, 2003.

[86] David Schönstein and B. Katz. HRTF selection for binaural synthesis from a database using morphological parameters. In *Proc. of the* 20^{th} *International Congress on Acoustics (ICA)*, Syndey, Australia, 2010.

[87] Dmitry N. Zotkin, Ramani Duraiswami, Larry S. Davis, Ankur Mohan, and Vikas Raykar. Virtual audio system customization using visual matching of ear parameters. In *Proc. of the* 16^{th} *International Conference on Pattern Recognition (ICPR)*, volume 3, pages 1003–1006, Quebec, Canada, 2002.

[88] Asta Kärkkäinen, Leo Kärkkäinen, and Tomi Huttunen. Practical procedure for large scale personalized head related transfer function acquisition. In *Proc of the* 51^{st} *Audio Engineering Society Conference (AES)*, Helsinki, Finland, 2013.

[89] M. Zhang, R. A. Kennedy, T. D. Abhayapala, and W. Zhang. Statistical method to identify key anthropometric parameters in HRTF individualization. In *Proc. of the Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, pages 213–218, 2011.

[90] Hongmei Hu, Lin Zhou, Jie Zhang, Hao Ma, and Zhenyang Wu. Head related transfer function personalization based on multiple regression analysis. In *Proc. of the International Conference on Computational Intelligence and Security (CIS)*, volume 2, pages 1829–1832, Guangzhou, China, 2006.

[91] K. Genuit. Ein modell zur beschreibung von außenohr übertragungseigenschaften. *Diss. RWTH Aachen*, 1984.

[92] Vikas C. Raykar, Ramani Duraiswami, and B. Yegnanarayana. Extracting the frequencies of the pinna spectral notches in measured head related impulse responses. *The Journal of the Acoustical Society of America*, 118(1):364–374, 2005.

[93] Simone Spagnol, Michele Geronazzo, and Federico Avanzini. Fitting pinna-related transfer functions to anthropometry for binaural sound rendering. In *Proc. of the IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pages 194–199, Saint Malo, France, 2010.

[94] Enrique A. Lopez-Poveda and Ray Meddis. A physical model of sound diffraction and reflections in the human concha. *The Journal of the Acoustical Society of America*, 100(5):3248–3259, 1996.

[95] Edgar A. G. Shaw. *Acoustical features of the human external ear*, pages 25–47. Lawrence Erlbaum Associates, Mahwah, NJ, USA, 1997.

[96] V. Ralph Algazi, Richard O. Duda, and Patrick Satarzadeh. Physical and filter pinna models based on anthropometry. In *Proc. of the* 122^{th} *Audio Engineering Society Convention (AES)*, Vienna, Austria, 2007.

[97] Carlos Avendano, V. Ralph Algazi, and Richard O. Duda. A head-and-torso model for low-frequency binaural elevation effects. In *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 179–182, New Paltz, NY, USA, 1999.

[98] V. Ralph Algazi, Richard O. Duda, and Dennis M. Thompson. The use of head-and-torso models for improved spatial sound synthesis. In *Proc. of the* 113^{th} *Audio Engineering Society Convention (AES)*, Los Angeles, CA, USA, 2002.

[99] V. Ralph Algazi, Carlos Avendano, and Richard O. Duda. Elevation localization and head-related transfer function analysis at low frequencies. *The Journal of the Acoustical Society of America*, 109:1110–1122, 2001.

[100] Philip Kortum. *HCI Beyond the GUI: Design for Haptic, Speech, Olfactory, and Other Nontraditional Interfaces*. Morgan Kaufmann Series in Interactive Technologies. Morgan Kaufmann, 2008.

[101] ITU-T recommendation M.1677-1 (10/09): International Morse code, 2009.

[102] Eoin Brazil, Mikael Fernstrom, and John Bowers. Exploring concurrent auditory icon recognition. In *Proc. of the* 15^{th} *International Conference on Auditory Display (ICAD)*, volume 9, Copenhagen, Denmark, 2009.

[103] William W. Gaver. The sonicfinder: An interface that uses auditory icons. *Human-Computer Interaction*, 4(1):67–94, 1989.

[104] Paul Robare and Jodi Forlizzi. Timelines sound in computing: a short history. *interactions*, 16(1):62–65, 2009.

[105] George Humphrey. The psychology of the gestalt. *Journal of Educational Psychology*, 15(7):401–412, 1924.

[106] Gemma Calvert, Charles Spence, and Barry E. Stein. *The handbook of multisensory processes*. The MIT Press, 2004.

[107] Jeremy I. Skipper, Virginie van Wassenhove, Howard C. Nusbaum, and Steven L. Small. Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex*, 17(10):2387–2399, 2007.

[108] Sharon Oviatt, Rachel Coulston, and Rebecca Lunsford. When do we interact multimodally? Cognitive load and multimodal communication patterns. In *Proc. of the* 6^{th} *international conference on Multimodal interfaces (ICMI)*, pages 129–136, Pittsburgh, USA, 2004.

[109] Natalie Ruiz, Ronnie Taib, and Fang Chen. Examining the redundancy of multimodal input. In *Proc. of the* 18^{th} *Australia conference on Computer-Human Interaction: Design: Activities, Artefacts and Environments (OZCHI)*, pages 389–392, Sydney, Australia, 2006.

[110] Ju-Hwan Lee and Charles Spence. Assessing the benefits of multimodal feedback on dual-task performance under demanding conditions. In *Proc. of the* 22^{nd} *British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction*, volume 1, pages 185–192, Liverpool, United Kingdom, 2008.

[111] Kurt A. Kaczmarek, John G. Webster, Paul Bach-y Rita, and Willis J. Tompkins. Electrotactile and vibrotactile displays for sensory substitution systems. *Biomedical Engineering, IEEE Transactions on*, 38(1):1–16, 1991.

[112] Alistair D. N. Edwards. Soundtrack: An auditory interface for blind users. *Human-Computer Interaction*, 4(1):45–66, 1989.

[113] Richard S. Schwerdtfeger. Making the GUI talk. *Byte Magazine*, (Dec.):118–128, 1991.

[114] Robert W. Massof. Auditory assistive devices for the blind. In *Proc. of the* 9^{th} *International Conference on Auditory Display (ICAD)*, pages 271–275, Boston, MA, USA, 2003.

[115] JAWS screen reading software (http://www.freedomscientific.com/products/fs/jaws-product-page.asp), 2013.

[116] GW Micro - window-eyes (http://www.gwmicro.com/window-eyes/), 2013.

[117] Dolphin - supernova (http://www.yourdolphin.co.uk/productdetail.asp?id=1), 2013.

[118] Stephen A. Brewster, Peter C. Wright, and Alistair D. N. Edwards. An evaluation of earcons for use in auditory human-computer interfaces. In *Proc. of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 222–227, Amsterdam, The Netherlands, 1993.

[119] J. Jay Todd, Daryl Fougnie, and René Marois. Visual short-term memory load suppresses temporo-parietal junction activity and induces inattentional blindness. *Psychological Science*, 16(12):965–972, 2005.

[120] Jack Loomis, Reginald Golledge, and Roberta Klatzky. Navigation system for the blind: Auditory display modes and guidance. *Presence*, 7:193–203, 1998.

[121] Shengdong Zhao, Pierre Dragicevic, Mark Chignell, Ravin Balakrishnan, and Patrick Baudisch. Earpod: eyes-free menu selection using touch input and reactive audio feedback. In *Proc. of of the SIGCHI conference on Human factors in computing systems (CHI '07)*, pages 1395–1404, San Jose, California, USA, 2007.

[122] Raine A. Kajastila and Tapio Lokki. A gesture-based and eyes-free control method for mobile devices. In *Proc. of CHI'09 Extended Abstracts on Human Factors in Computing Systems*, pages 3559–3564, Boston, MA, USA, 2009.

[123] Thomas Hermann, Andy Hunt, and John G. Neuhoff. *The Sonification Handbook*. Logos Verlag, 2011.

[124] Albert S. Bregman. *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[125] Bernd Meyer. *Perception of speech and sound*, pages 61–82. Springer, 2008.

[126] Hugo Fastl and Eberhard Zwicker. *Psychoacoustics: facts and models*. Springer, 3rd ed. edition, 2007.

[127] Ward R. Drennan, Stuart Gatehouse, and Catherine Lever. Perceptual segregation of competing speech sounds: the role of spatial location. *The Journal of the Acoustical Society of America*, 114:2178–2189, 2003.

[128] Donald Eric Broadbent. The role of auditory localization in attention and memory span. *Journal of experimental psychology*, 47(3):191–196, 1954.

[129] Donald D. Dirks and Richard H. Wilson. The effect of spatially separated sound sources on speech intelligibility. *Journal of Speech, Language and Hearing Research*, 12(1):5–38, 1969.

[130] Yoshiro Miyata and Donald A. Norman. Psychological issues in support of multiple activities. *User centered system design: New perspectives on human-computer interaction*, pages 265–284, 1986.

[131] Martijn Schreuder, Benjamin Blankertz, and Michael Tangermann. A new auditory multi-class brain-computer interface paradigm: spatial hearing as an informative cue. *PLoS One*, 5(4), 2010.

[132] Donald D. Greenwood. Auditory masking and the critical band. *The Journal of the Acoustical Society of America*, 33(4):484–502, 1961.

[133] Thomas Christensen. *The Cambridge history of Western music theory*, volume 3. Cambridge University Press, 2002.

[134] J. S. Bradley and H. Sato. Speech intelligibility test results for grades 1, 3 and 6 children in real classrooms. In *Proc. of the* 18^{th} *International Congress on Acoustics (ICA)*, pages 1–4, Kyoto, Japan, 2004.

[135] Kevin S. LaBar and Roberto Cabeza. Cognitive neuroscience of emotional memory. *Nature Reviews Neuroscience*, 7(1):54–64, 2006.

[136] Irwin Pollack and J. M. Pickett. Cocktail party effect. *The Journal of the Acoustical Society of America*, 29:1262, 1957.

[137] D. E. Broadbent. *Perception and communication*. Pergamon, Oxford, 1958.

[138] Irvin Rock and Stephen Palmer. The legacy of gestalt psychology. *Scientific American*, 263(6):84–90, 1990.

[139] George A. Miller. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97, 1956.

[140] Fred Paas and Jeroen J.G. Van Merriënboer. Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review*, 6(4):351–371, 1994.

[141] Fred Paas, Juhani E. Tuovinen, Huib Tabbers, and Pascal W. M. Van Gerven. Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist*, 38(1):63–71, 2003.

[142] Sandra G. Hart and Lowell E. Staveland. Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Human mental workload*, 1(3):139–183, 1988.

[143] Paul A. Kirschner. Cognitive load theory: Implications of cognitive load theory on the design of learning. *Learning and instruction*, 12(1):1–10, 2002.

[144] Bruce N. Walker and Gregory Kramer. Mappings and metaphors in auditory displays: An experimental assessment. *ACM Transactions on Applied Perception (TAP)*, 2(4):407–412, 2005.

[145] Gregory Kramer. *Auditory display: Sonification, audification, and auditory interfaces*. Addison-Wesley Reading, MA, 1994.

[146] Mark Ballora, Bruce Pennycook, Plamen C. Ivanov, Leon Glass, and Ary L. Goldberger. Heart rate sonification: A new approach to medical diagnosis. *Leonardo*, 37(1):41–46, 2004.

[147] Petr Janata and Edward Childs. Marketbuzz: Sonification of real-time financial data. In *Proc. of the* 10^{th} *International Conference of Auditory Display (ICAD)*, Sydney, Australia, 2004.

[148] Samuel Van Ransbeeck and Carlos Guedes. Stockwatch, a tool for composition with complex data. *Parsons Journal for Information Mapping*, 1(3), 2009.

[149] Thomas Hermann. Taxonomy and definitions for sonification and auditory display. In *Proc. of the* 14^{th} *International Conference on Auditory Display (ICAD)*, Paris, France, 2008.

[150] John Williamson, Roderick Murray-Smith, and Stephen Hughes. Shoogle: excitatory multimodal interaction on mobile devices. In *Proc. of the SIGCHI conference on Human factors in computing systems (CHI '07)*, pages 121–124, San Jose, CA, USA, 2007.

[151] William W. Gaver. Auditory icons: Using sound in computer interfaces. *Human-Computer Interaction*, 2(2):167–177, 1986.

[152] James L. Alty, Dimitrios Rigas, and Paul Vickers. Music and speech in auditory interfaces: When is one mode more appropriate than another? In *Proc. of the* 11^{th} *International Conference on Auditory Display (ICAD)*, pages 351–357, Limerick, Ireland, 2005.

[153] Richard W. Sproat and Joseph P. Olive. Text-to-speech synthesis. *AT & T technical journal*, 74(2):35–44, 1995.

[154] Thierry Dutoit. *An introduction to text to speech synthesis*, volume 3. Springer, 1997.

[155] Uwe D. Reichel and Hartmut R. Pfitzinger. Text preprocessing for speech synthesis. *Department of Phonetics and Speech Communication University of Munich*, 2006.

[156] A. Chauhan, V. Chauhan, G. Singh, C. Choudhary, and P. Arya. Design and development of a text-to-speech synthesizer system,. *International Journal on Electronics & Communication Technology*, 2(3):42–45, 2011.

[157] Amy T. Neel. Formant detail needed for vowel identification. *Acoustics Research Letters Online*, 5(4):125–131, 2004.

[158] Thomas Styger and Éric Keller. *Formant synthesis*, pages 109–128. John Wiley and Sons Ltd., 1995.

[159] Bruce N. Walker, Amanda Nance, and Jeffrey Lindsay. Spearcons: Speech-based earcons improve navigation performance in auditory menus. In *Proc. of the* 12^{th} *International Conference on Auditory Display*, pages 63–68, London, UK, 2006.

[160] Dianne K. Palladino and Bruce N. Walker. Navigation efficiency of two dimensional auditory menus using spearcon enhancements. In *Proc. of the Human Factors and Ergonomics Society* 52^{nd} *Annual Meeting*, pages 1262–1266, New York, NY, USA, 2008.

[161] Dianne K. Palladino. Efficiency of spearcon-enhanced navigation of one dimensional electronic menus. 2008.

[162] Dianne K. Palladino and Bruce N. Walker. Learning rates for auditory menus enhanced with spearcons versus earcons. In *Proc. of the* 13^{th} *international conference on auditory display (ICAD)*, pages 274–279, 2007.

[163] Tilman Dingler, Jeffrey Lindsay, and Bruce N. Walker. Learnability of sound cues for environmental features: Auditory icons, earcons, spearcons, and speech. In *Proc. of the* 14^{th} *International Conference on Auditory Display (ICAD)*, pages 1–6, Paris, France, 2008.

[164] Pavani Yalla and Bruce N. Walker. Advanced auditory menus. *Georgia Institute of Technology GVU Center GIT-GVU-07-12*, 2007.

[165] David K. McGookin and Stephen A. Brewster. An investigation into the identification of concurrently presented earcons. In *Proc. of the* 9^{th} *international conference on auditory display (ICAD)*, pages 42–46, Boston, MA, USA, 2003.

[166] David K. McGookin and Stephen A. Brewster. Advantages and issues with concurrent audio presentation as part of an auditory display. In *Proc. of the* 12^{th} *International Conference on Auditory Display (ICAD)*, pages 40–55, London, United Kingdom, 2006.

[167] Jaka Sodnik, Grega Jakus, and Sašo Tomažič. Multiple spatial sounds in hierarchical menu navigation for visually impaired computer users. *International journal of human-computer studies*, 69(1):100–112, 2011.

[168] David K. McGookin and Stephen A. Brewster. Understanding concurrent earcons: Applying auditory scene analysis principles to concurrent earcon recognition. *ACM Transactions on Applied Perception (TAP)*, 1(2):130–155, 2004.

[169] Hong Jun Song and Kirsty Beilharz. Concurrent auditory stream discrimination in auditory graphing. *Journal of Computers*, 3:79–87, 2007.

[170] Mark R. Anderson, Durand Begault, Martine Godfroy, Joel D. Miller, and Elizabeth M. Wenzel. Applying spatial audio to human interfaces: 25 years of NASA experience. In *Proc. of the* 40^{th} *International Audio Engineering Society Conference (AES)*, Tokyo, Japan, 2010.

[171] Kai Crispien, Klaus Fellbaum, Anthony Savidis, and Constantine Stephanidis. A 3D-auditory environment for hierarchical navigation in non-visual interaction. In *Proc. of the* 3^{rd} *international conference on auditory display (ICAD)*, Palo Alto, CA, USA, 1996.

[172] B.W. Anderson and J.T. Kalb. English verification of the sti method for estimating speech intelligiblity of a communications channel. *J. Acoust. Soc. Am.*, 81:1982–1985, 1987.

[173] Hiroshi Ishii, Craig Wisneski, Scott Brave, Andrew Dahley, Matt Gorbet, Brygg Ullmer, and Paul Yarin. ambientROOM: integrating ambient media with architectural space. In *Proc. of the ACM CHI 98 Human Factors in Computing Systems Conference*, pages 173–174, Los Angeles, CA, USA, 1998.

[174] Tara Matthews, Tye Rattenbury, Scott Carter, Anind Dey, and Jennifer Mankoff. A peripheral display toolkit. Technical report, EECS Department, University of California, Berkeley, 2003.

[175] John Stasko, Todd Miller, Zachary Pousman, Christopher Plaue, and Osman Ullah. *Personalized peripheral information awareness through information art*, pages 18–35. Springer, 2004.

[176] D. Scott McCrickard, Christa M. Chewar, Jacob P. Somervell, and Ali Ndiwalana. A model for notification systems evaluation-assessing user goals for multitasking activity. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 10(4):312–338, 2003.

[177] Zachary Pousman and John Stasko. A taxonomy of ambient information systems: four patterns of design. In *Proc. of the working conference on Advanced visual interfaces (AVI '06)*, pages 67–74, Venice, Italy, 2006.

[178] Craig Wisneski, Hiroshi Ishii, Andrew Dahley, Matt Gorbet, Scott Brave, Brygg Ullmer, and Paul Yarin. *Ambient displays: Turning architectural space into an interface between people and digital information*, pages 22–32. Springer, 1998.

[179] Jonathan Cohen. "Kirk here:" using genre sounds to monitor background activity. In *Proc. of the INTERACT'93 and CHI'93 Conference Companion on Human Factors in Computing Systems*, pages 63–64, Amsterdam, The Netherlands, 1993.

[180] Fredrik Kilander and Peter Lönnqvist. A weakly intrusive ambient soundscape for intuitive state perception. *Continuity in future computing systems*, pages 70–74, 2001.

[181] Fredrik Kilander and Peter Lönnqvist. A whisper in the woods - an ambient soundscape for peripheral awareness of remote processes. In *Proc. of the* 8^{th} *International Conference on Auditory Display (ICAD)*, Kyoto, Japan, 2002.

[182] Eoin Brazil and J. M. Fernström. Investigating ambient auditory information systems. In *Proc. of the* 13^{th} *International Conference on Auditory Display (ICAD)*, pages 326–333, Montreal, 2007.

[183] Daniel Lee Ashbrook. *Enabling mobile microinteractions*. PhD thesis, 2010.

[184] Dan Morris, T. Scott Saponas, and Desney Tan. Emerging input technologies for always-available mobile interaction. *Foundations and Trends in Human-Computer Interaction*, 4(4):245–316, 2010.

[185] Sami Ronkainen, Jonna Häkkilä, Saana Kaleva, Ashley Colley, and Jukka Linjama. Tap input as an embedded interaction method for mobile devices. In *Proc. of the* 1^{st} *International conference on Tangible and embedded interaction (TEI'07)*, pages 263–270, Baton Rouge, LA, USA, 2007.

[186] Scott E. Hudson, Chris Harrison, Beverly L. Harrison, and Anthony LaMarca. Whack gestures: inexact and inattentive interaction with mobile devices. In *Proc. of the* 4^{th} *international conference on Tangible, embedded, and embodied interaction (TEI'10)*, pages 109–112, Cambridge, MA, USA, 2010.

[187] T. Scott Saponas, Chris Harrison, and Hrvoje Benko. PocketTouch: Through-fabric capacitive touch input. In *Proc. of the* 24^{th} *annual ACM symposium on User interface software and technology (UIST'11)*, pages 303–308, Santa Barbara, CA, USA, 2011.

[188] Liwei Chan, Rong-Hao Liang, Ming-Chang Tsai, Kai-Yin Cheng, Chao-Huai Su, Mike Y. Chen, Wen-Huang Cheng, and Bing-Yu Chen. FingerPad: private and subtle interaction using fingertips. In *Proc. of the* 26^{th} *annual ACM symposium on User interface software and technology (UIST'13)*, pages 255–260, St Andrews, United Kingdom, 2013.

[189] Xing-Dong Yang, Tovi Grossman, Daniel Wigdor, and George Fitzmaurice. Magic finger: always-available input through finger instrumentation. In *Proc. of the* 25^{th} *annual ACM symposium on User interface software and technology (UIST'12)*, pages 147–156, Cambridge, MA, USA, 2012.

[190] Roderick Murray-Smith, John Williamson, Stephen Hughes, and Torben Quaade. Stane: synthesized surfaces for tactile input. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*, pages 1299–1302, Florence, Italy, 2008.

[191] Julius O. Smith III. *Introduction to digital filters: with audio applications*, volume 2. W3K Publishing, 2007.

[192] K. A. Morris. What is hysteresis? *Applied Mechanics Reviews*, 64(5), 2010. 10.1115/1.4007112.

[193] Willie Walker, Paul Lamere, and Philip Kwok. FreeTTS 1.2 ((http://www.freetts.sourceforge.net).

[194] Michael N. Tombu, Christopher L. Asplund, Paul E. Dux, Douglass Godwin, Justin W. Martin, and René Marois. A unified attentional bottleneck in the human brain. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 108(33):13426–13431, 2011.

[195] Christine Rosen. The myth of multitasking. *The New Atlantis*, 20:105–110, 2008.

[196] Joshua S. Rubinstein, David E. Meyer, and Jeffrey E. Evans. Executive control of cognitive processes in task switching. *Journal of Experimental Psychology: Human Perception and Performance*, 27(4):763–797, 2001.

[197] Barry Arons. A review of the cocktail party effect. *Journal of the American Voice I/O Society*, 12(7):35–50, 1992.

[198] William A. Yost, Raymond H. Dye, and Stanley Sheft. A simulated "cocktail party" with up to three sound sources. *Perception & Psychophysics*, 58(7):1026–1036, 1996.

[199] Kai Crispien and Tasso Ehrenberg. Evaluation of the-cocktail-party effect-for multiple speech stimuli within a spatial auditory display. *Journal of the Audio Engineering Society*, 43(11):932–941, 1995.

[200] Lisa J. Stifelman. The cocktail party effect in auditory interfaces: A study of simultaneous presentation. Technical report, MIT Media Laboratory, 1994.

[201] Bobby Owsinski and Malcolm O'Brien. *The mixing engineer's handbook*. Thomson Course Technology, 2006.

[202] Eric R. Spangenberg, Kevin E. Voss, and Ayn E. Crowley. Measuring the hedonic and utilitarian dimensions of attitudes: a generally applicable scale. *Advances in Consumer Research*, 24:235–241, 1997.

[203] Robert S. Bolia, W. Todd Nelson, Mark A. Ericson, and Brian D. Simpson. A speech corpus for multitalker communications research. *The Journal of the Acoustical Society of America*, 107(2):1065–1066, 2000.

[204] Douglas S. Brungart. Evaluation of speech intelligibility with the coordinate response measure. *The Journal of the Acoustical Society of America*, 109:2276–2279, 2001.

[205] Richard L. McKinley and Mark A. Ericson. Flight demonstration of a 3-d auditory display. *Binaural and Spatial Hearing in Real and Virtual Environments*, pages 683–699, 1997.

[206] Doug A. Bowman, Ernst Kruijff, Joseph J. LaViola Jr, and Ivan Poupyrev. *3D User Interfaces: Theory and Practice*. Addison-Wesley, 1^{st} ed. edition, 2004.

[207] Kai Crispien and Helen Petrie. Providing access to guis for blind people using a multimedia system based on spatial audio presentation. *PREPRINTS-AUDIO ENGINEERING SOCIETY*, 1993.

[208] Niklas Røber and Maic Masuch. Interacting with sound: An interaction paradigm for virtual auditory worlds. In *Proc. of the* 10^{th} *Meeting of the International Conference on Auditory Display (ICAD)*, Sydney, Australia, 2004.

[209] Hilko Donker, Palle Klante, and Peter Gorny. The design of auditory user interfaces for blind users. In *Proc. of the* 2^{nd} *Nordic conference on Human-computer interaction (NordiCHI'02)*, pages 149–156, Aarhus, Denmark, 2002.

[210] E. H. Dooijes and J. D. Mulder. Spatial audio in graphical applications. Technical report, CWI Netherlands, 1994.

[211] Ian J. Pitt and Allstair D. N. Edwards. Pointing in an auditory interface for blind users. In *Proc. of the IEEE International Conference on Systems, Man and Cybernetics*, volume 1, pages 280–285, Vancouver, BC, Canada, 1995.

[212] Sebastian Deterding, Miguel Sicart, Lennart Nacke, Kenton O'Hara, and Dan Dixon. Gamification. using game-design elements in non-gaming contexts. In *Proceedings of the annual conference extended abstracts on Human factors in computing systems (CHI EA '11)*, pages 2425–2428, Vancouver, BC, Canada, 2011.

[213] Thomas Svedström and Aki Härmä. Eyes-free interaction for personal media devices. In *Proc. of the* 136^{th} *Audio Engineering Society Convention (AES)*, Berlin, Germany, 2014.

[214] Aki Härmä, Ralph van Dinther, Thomas Svedström, Munhum Park, and Jeroen Koppens. Personalization of headphone spatialization based on the relative localization error in an auditory gaming interface. In *Proc. of the* 132^{nd} *International Audio Engineering Society Convention (AES)*, Budapest, Hungary, 2012.

Appendix A

Consumer insights

1. When I am on-the-go I usually listen to music with my personal device (PMD). My PMD is typically online and I get announcements like messages from my friends or calendar indications. At times this gets quite messy when there is a lot going on. Therefore, my music listening suffers from all alarms and indicator messages, or I may miss some important signals because of the music I am listening. I wish there was a meaningful way to combine entertainment and messages into a smooth and pleasant experience.

2. I am using my PD for staying in touch with my friends. We often use text/image-based interfaces (SMS, twitter, facebook). I often feel like answering immediately to short messages. Often it would be enough to just give "yes/no/maybe" as a reply. I can do that in 10-20 seconds with my PD. However, this requires that I have full attention to the device and may lead to dangerous traffic situations. I wish I could do simple real-time messaging (sms/twitter/facebook) without looking at the screen.

3. I listen to the music with my portable player. I often update my player with new contents and then I often like to listen to browse and explore my new collection when I am *on-the-go*. Changing from one album to another is basically simple but I need to see the navigation screen in order to choose the album I want. It is very dangerous to do that in a busy traffic (e.g., on a bicycle) and therefore I would really appreciate a solution which helps me navigate simple menus without eyes.

4. I use my PD for navigational purposes (GPS) when I go to places where I have not previously been. I want to get to my destination, but using the map and simultaneously looking all the exciting new things is

sometimes annoying. After all, I am not so good with maps. I would love to have someone showing or telling me which direction to go, but I really do not like the way navigators instruct me: "Turn right after 200 meters". I want to decide my own path and stay on the map.

5. I like to know what is going in the world and in my local area. Sometimes I just do not have the time to watch the news or read the newspaper. Previously I was listening to radio a lot and besides to the music, I could also hear all the latest news and weather forecasts. Nowadays I use my PD for music listening. I wish that I could get the latest news also while listening to my favorite collection of music.

Appendix B

Auditory menu structure

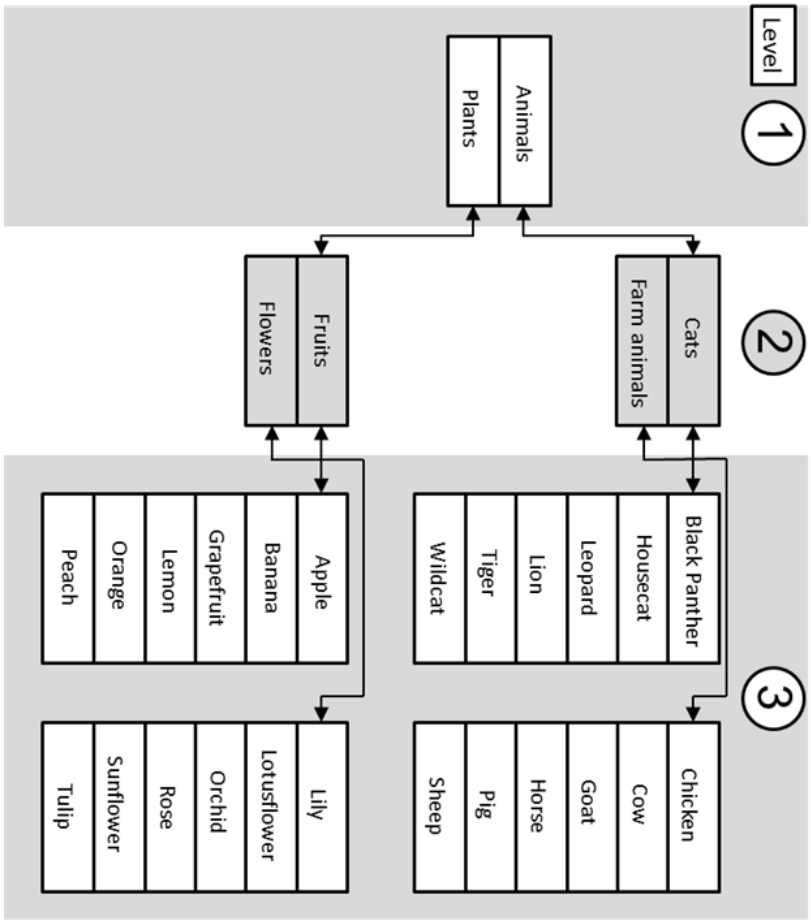

Figure B.1: The menu structure used in chapter 4.

Appendix C

GUI for the reaction and menu browsing time experiment

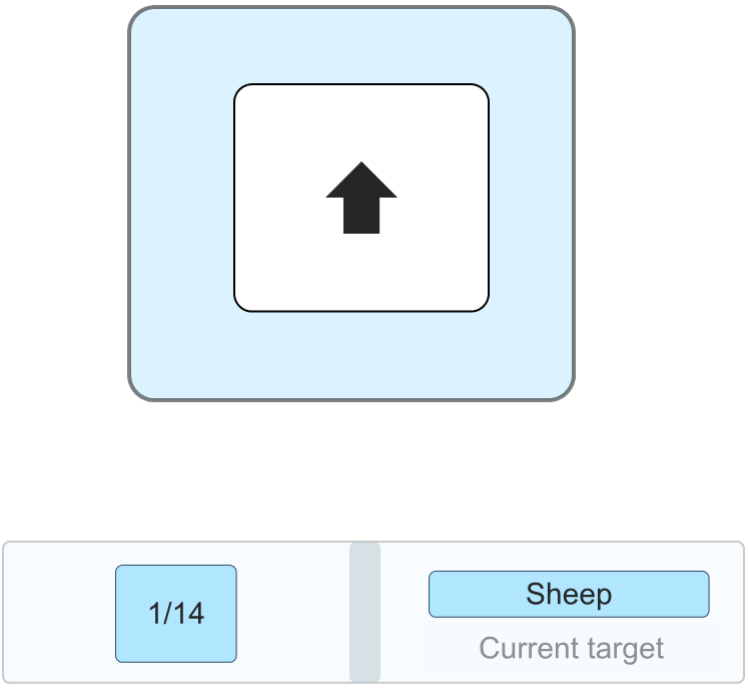

Figure C.1: The GUI for chapter 4. It was used to present the current target and to display the arrows to which the user had to react.

Appendix D

GUI for the multilayer auditory interface experiment

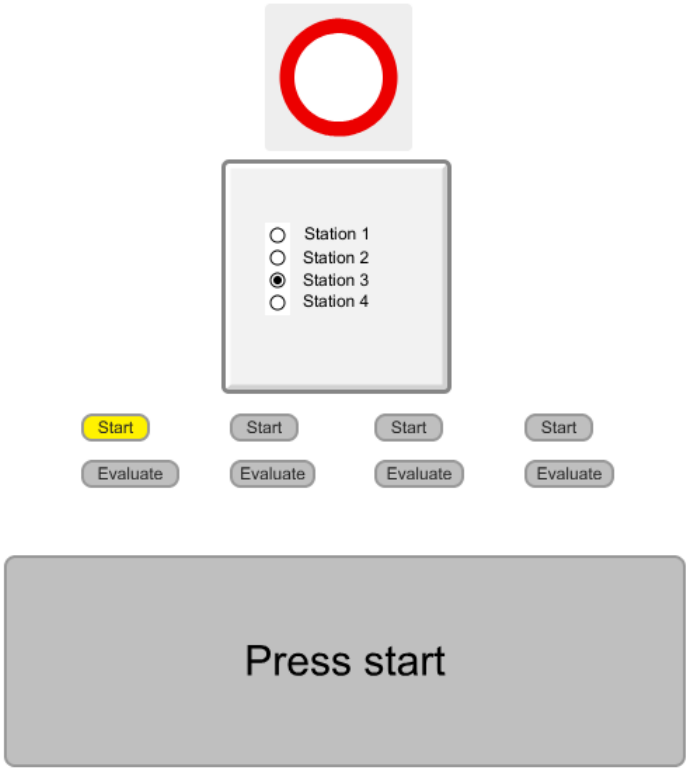

Figure D.1: The GUI that was used in the multilayer auditory interface experiment in chapter 5.

Appendix E

GUI for the HRTF personalization experiment

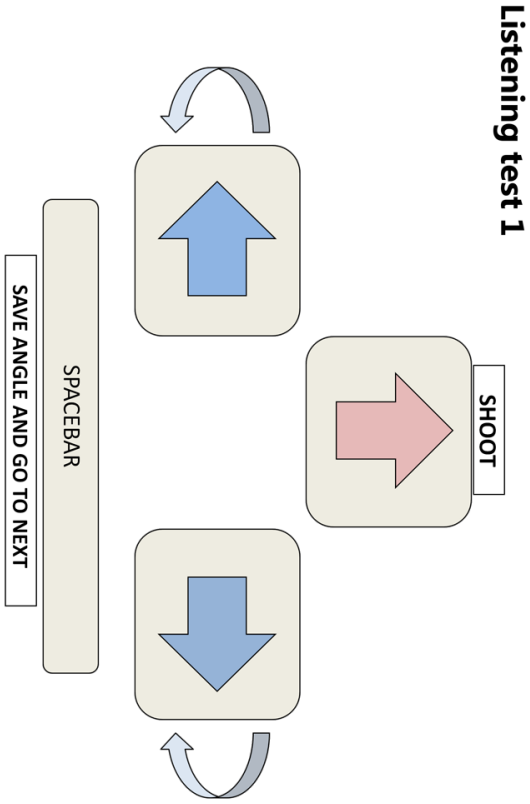

Figure E.1: The GUI for the HRTF personalization test in chapter 6.